RAND

Updating the Classification System for the Medicare Inpatient Rehabilitation Prospective Payment System

Grace M. Carter, Daniel A. Relles, and Gregory K. Ridgeway

DRU-2491-HCFA

February 2001

RAND Health

PREFACE

This draft is a preliminary version of a report on the construction of an updated set of Function Independence Measure-Function Related Groups (FIM-FRGs, or just FRGs). It was written for a project in support of the Health Care Financing Administration's (HCFA) design, development, implementation, monitoring, and refining of a Prospective Payment System (PPS) for inpatient rehabilitation. Such an inpatient rehabilitation facility PPS (IRF PPS) was mandated in the Balanced Budget Act of 1997. This report is being circulated in this form in order to obtain comments from a technical expert panel. After comments have been received, the report will be revised and made publicly available.

The research reported here was supported by HCFA through contract 500-95-0056 with RAND.

CONTENTS

TABLES

FIGURES

EXECUTIVE SUMMARY

In the Balanced Budget Act of 1997, Congress mandated that the Health Care Financing Administration (HCFA) implement a Prospective Payment System (PPS) for inpatient rehabilitation. RAND contracted with HCFA to carry out the research, and recruited a Technical Expert Panel (TEP) to advise on issues related to the design and development of such a PPS. The TEP convened in May 2000 to review our Interim Report. A key topic discussed there was the construction of Function Related Groups (FRGs), which will be the basis of the payment classification system. This report follows up the TEP suggestions, both formal and informal, for further research into FRGs.

In the interim report, we used CART to obtain candidate FRGs. The TEP asked for a broader context within which to view the results. Specifically, the TEP wanted us to

(1) Explore alternative model forms. Develop models to compete with CART in terms of having strong predictive performance.

(2) Consider indices of function in addition to the cognitive and motor scores. Payment formulas based on these measures might offer better estimates of cost.

(3) Evaluate out-of-sample performance of the models. An important element of a payment system is whether payment formulas developed from data in one year apply in future years.

In this report, we describe the steps we have taken to update FRGs on newer data while incorporating the above suggestions.

Data

We used hospital cost reports, discharge abstracts from MEDPAR, and functional independence measure (FIM) data for Medicare discharges in years 1996 through 1999. We added 1998 and 1999 data to our database since the TEP meeting. We use the departmental method to estimate the accounting cost of each case in the MEDPAR sample. The FIM data come from the Uniform Data System for medical rehabilitation (UDSmr), from the Clinical Outcomes Systems (COS) data for medical rehabilitation

(1996 and 1997), and from Health South Rehabilitation Hospitals (1998 and 1999). Our sample covers about half of all inpatient rehabilitation facility Medicare patients in the first two years, but 70 percent of this population by 1999.

Modeling Methods

We identify the three basic suggestions of the TEP and describe a computational experiment to carry them out, leading to the identification of specific methods for determining and evaluating FRGs. The computational experiment varied six types of models over six types of indices over four years of data. The dependent variable in all the models was the logarithm of wage adjusted cost.

(1) Explore alternative model forms. CART is the traditional method of generating FRGs and a reasonable method for determining rules to classify patients into groups that explain cost. CART is efficient at producing simple and effective rules for prediction, but it also has its limitations. In particular, it adheres to a particular functional form, and its fitting algorithm does not necessarily produce a global optimum. So we compared its performance with other methods that are known in the statistics literature: ordinary linear least squares (OLS), generalized additive models (GAM), and multiple adaptive regression trees (MART). These were in addition to three variations of CART that differed in their stopping rules. To determine which models fit best, we assessed each model's out-of-sample predictive performance. Our goal was to determine what percent of the "predictable" variation in costs CART could predict.

(2) Explore the predictive ability of other functional measures. The 13-item FIM motor scale and the five-item FIM cognitive scale are well-established measures of motor and cognitive ability. We examined individual FIM items to see if each one entered in the expected direction--if not, it would suggest problems with their presence in the scale and we would consider removing them. We also experimented with several of the sub-

scales described in Stineman, Jette, et al. (1997). These split out the standard motor index into dimensions that pertain to different body areas and types of function.

(3) Examine the stability of predictions across years. Our previous results were based on 1996 and 1997 data, and did not give us much latitude for examining stability over time. With the addition of 1998 and 1999 data, we have the option of fitting models within each year and seeing how well they do on other years. We also have the ability to pool multiple years worth of data for RICs that are small and hence have imprecise estimates of cost. Table S.1 illustrates the combinations of fitting and evaluation years that we used.

Table S.1

Combination of Fitting and Evaluation Periods Examined

	Evaluation Period			
Fitting Period	**1996**	**1997**	**1998**	**1999**
1996	.	x	x	x
1997	x	.	x	x
1998	x	x	.	x
1999	x	x	x	.
1996-97	.	.	x	x
1998-99	x	x	.	.

Results--Item Level Analysis

We regressed log cost (OLS and GAM) on the responses to individual FIM items--eating, walking, etc. We wanted to know whether the individual items appeared to influence costs in the expected direction: higher FIM scores should mean lower costs, so coefficients should be negative. Randomness alone would produce numerous positive regression coefficients, so we looked for items with consistently positive coefficients across all four years of data and many RICs. The unmistakable pattern is that both tub transfers and comprehension often have the wrong sign in OLS (and GAM) regressions--costs were higher when the functional independence measure was higher.

The response to the transfer to tub/shower depends on the situation being tested—whether tub or shower and whether an assistive device is

used. Thus it does not provide a measure of the relative capability of different patients. We have no similar rationale for the comprehension results.

Based on these results, we removed transfer to tub from all indices that included it, and we eliminated comprehension from the cognitive index in order to compare the resulting cost predictions with those from the standard motor and cognitive scores. Table S.2 indicates the combinations of indices we selected for further investigation.

Table S.2

The Candidate Indices

<table>
<tr><th>Items</th><th>M13C5</th><th>M12C5</th><th>M12C4</th><th>StJe3</th><th>StJe5</th></tr>
<tr><td>transfer to tub/shower</td><td rowspan="12">standard motor</td><td>X</td><td>X</td><td>X</td><td>X</td></tr>
<tr><td>transfer to bed/chair</td><td rowspan="12">motor excluding trftub</td><td rowspan="12">motor excluding trftub</td><td rowspan="4">mobility excluding trftub</td><td rowspan="2">transfer excluding trftub</td></tr>
<tr><td>transfer to toilet</td></tr>
<tr><td>Walking/ wheelchair</td><td rowspan="2">locomotion</td></tr>
<tr><td>stairs</td></tr>
<tr><td>bladder</td><td rowspan="6">ADLS</td><td>sphincter</td></tr>
<tr><td>bowel</td></tr>
<tr><td>eating</td><td rowspan="5">self care</td></tr>
<tr><td>grooming</td></tr>
<tr><td>bathing</td></tr>
<tr><td>dress upper</td></tr>
<tr><td>dress lower</td></tr>
<tr><td>toilet</td></tr>
<tr><td>comprehension</td><td rowspan="6">standard cognitive</td><td rowspan="6">standard cognitive</td><td>X</td><td rowspan="6">standard cognitive</td><td rowspan="6">standard cognitive</td></tr>
<tr><td>expression</td><td rowspan="5">cognitive excluding compreh</td></tr>
<tr><td>social interaction</td></tr>
<tr><td>problem solving</td></tr>
<tr><td>memory</td></tr>
</table>

Note: transfer to tub has been a traditional component of all these mobility indices. However, for reasons developed in Section 3.3.1, we take transfer to tub out of the relevant indices when the time comes to use them.

Results--Selecting a Gold Standard Model

A gold standard model can help us evaluate how well CART is doing--it gives us a measure of attainable residual standard deviation to

compare to the residual standard deviation we get from CART. It also will enable us in a later simulation exercise to assess the prediction bias for various combinations of demographic and hospital characteristics. These simulations will be reported in the project's final report.

MART and GAM are the candidates for gold standard status. Both are extremely flexible and can trace out prediction formulas with arbitrary shapes--not just linear shapes (as in OLS), not just step function shapes (as in CART). MART models allow interactions, GAM models are additive. So, MART is more flexible, but GAM fits are generally easier to interpret. Knowing that CART would not produce reasonable models with the individual FIM item scores, we chose not to work further with items at this point. We used all of the five remaining index sets and each 6 fitting periods described in the previous tables. We looked at out-of-sample root mean squared prediction error (RMSE) as a measure of quality of fits, both aggregated across RICs and disaggregated within RICs, and drew the following conclusions.

(1) MART and GAM do about equally well.

(2) The motor score without transfer to tub predicts costs slightly better than the standard motor score; eliminating comprehension from the cognitive score produces a further slight improvement in prediction accuracy in some cases.

3) Both GAM and MART seem to be able to make use of the sub-scales of the motor scale. RMSE goes down as the number of sub-scales goes up, and the RMSE is lowest for the most disaggregated set of indices StJe5.

(4) The RMSEs are all large, even for the best performing index set StJe5. About 15 percent of the within-RIC standard deviation, or 25 percent of the variance, is explainable. This translates to predicting about 38 percent of the total variance in cost, including the variance across RICs. But we cannot do better than that--case level costs are inherently unpredictable.

We decided to use MART with index set StJe5 as the gold standard. Prior to reviewing exactly which models to use within each RIC for our final report simulations, we assume that this model will provide a good

estimate of the percent of the explainable standard deviation attained by our CART models.

Results--Cost Patterns

We wanted to understand the marginal contribution of motor and cognitive scores to the estimated log cost. OLS coefficients provide such marginal estimates, but they enforce linear effects. GAM provides marginal estimates and allows arbitrary curvature. We attempted to understand the patterns of fit by graphing our GAM-M12C5 fits versus the motor and cognitive scales. Because the GAM fits were almost as good as MART's, we thought this would give an accurate portrayal of the cost versus scale relationships.

We found that the patterns of variation are described by a strong relationship between motor and cost--higher motor scores lower cost, and a weak relationship between cognitive and cost. The fitted curves do not appear to be far from monotone approximations that enforce an inverse relationship between cost and FIM scores. This implies that the data will support a "monotone" payment scheme where higher FIM scores never lead to higher payments, perhaps a politically desirable situation.

Results--Evaluating CART Models

The design criteria of the payment system require developing simple, understandable patient classification groups. CART is the ideal tool for building classification models. We considered three basic variations of stopping rules: (1) XVAL--the standard cross-validation method, which stops when CART thinks the minimum prediction error is achieved; (2) 1SD--the one standard deviation rule, which stops when CART thinks the prediction error is within one standard deviation of the minimum; and (3) INT--the number of nodes in the interim report (DRU-2309-HCFA, July 2000). We use two basic criteria to evaluate the alternatives: RMSE and parsimony.

In CART, the index with transfer to tub (M13C5) does noticeably worse than the index without this item in many years and stopping rules and never does substantially better. This is similar to our findings

with GAM and MART. The number of FRGs in the models is relatively similar.

If we drop comprehension from the cognitive index, we get a further slightly better prediction in some years and stopping rules. We discuss reasons in the text why one might want to include or exclude comprehension from the cognitive index, but we hope to get the TEP's input to help resolve this issue.

At the same number of nodes, either of the models based on sub-scale have substantially worse performance than any version of the motor and cognitive scales. With either the 1SD stopping rule or the XVAL rule, we get many more nodes. For example, with 1SD-CART in 1999, we get 50 percent more nodes than with the motor and cognitive scales (M12C5), with almost no improvement in RMSE.

We found that the CART 1 standard deviation rule produced fits that explained about 84 percent of the explainable standard deviation. We also found that the CART 1 standard deviation rule produced less than half the nodes of the cross-validation rule.

Results--Recommendations

Our tentative recommendation to HCFA is to use the CART model with the motor score without the transfer to tub/shower item and with the 1SD-stopping rule as the basis of the case mix groups. We drop transfer to tub/shower because we believe this item does not measure an absolute level of function and slightly decreases our ability to predict cost. We use two years of data in several small RICs. We are requesting CART input on several aspects of this recommendation.

After deciding to develop FRGs through 1SD-CART models using the index M12C5, we enforced monotonicity in the motor and cognitive scale and joined adjacent nodes of a tree where the cost estimates were similar. Our proposed model has 95 FRGs, with splits mostly on motor scores, but with age and cognitive scores playing a limited role. Age matters only at low motor scores and, where age matters, younger patients are more expensive. Cognitive function helps to define groups for patients with high motor scores in stroke, osteoarthritis, and in three of the brain injury, and spinal cord injury RICs, but splits

patients with low motor score in the joint replacement RIC. Lower cognitive cost predicts larger costs.

Using the gold standard models, we can assess how well the 95-node model is doing in out-of-sample years: it explains about 80 percent of the explainable standard deviation (81 percent of the variance). The payment system, of course, also exploits the variance across RICs in cost. About 34 percent of the total variance in the wage adjusted cost of cases discharged to the community is predicted by the proposed FRG system (38 percent by our gold standard models).

The final trees differ in some respects from the trees produced in the interim report. This is not surprising--CART is trying to fit step functions to continuous curves, so the cut-points are imprecisely determined. We think the important question is not whether the trees are identical but instead whether the tree models produce a consistent and accurate set of predictions. For now, we simply ask if one used the current FRGs and associated predictions, how different are these predictions over the different years? The plots in Section 4 demonstrate that this set of FRG models fits the data pretty well in all years. In future work, we will determine rules for updating the FRGs that result in fewer changes to the FRG definitions than would an annual refitting of the CART model.

ACKNOWLEDGMENTS

We thank our HCFA project officer, Carolyn Rimes, for her continued support throughout the project and for her rapid response to our requests for data. She also arranged frequent, very helpful telephone conversations with various HCFA staff. We would particularly like to thank Nora Hoban, Robert Kuhl, and Laurie Feinberg for their willing participation in these calls which helped us understand HCFA's analyses needs and HCFA's data.

We also thank Richard Linn, Director of UDSmr and the Center for Functional Assessment Research, State University of New York, and Jean Davis, Inpatient Director of Operations, Health South Hospitals, for the acquisition and use of their data and for their help in data interpretation.

1. INTRODUCTION

This report to the Technical Expert Panel (TEP) follows up comments made at the TEP meeting in May 2000. A presentation and discussion of FRGs at that meeting had brought forward three basic suggestions:

(1) Explore alternative model forms. Develop models to compete with CART in terms of having strong predictive performance. CART is limited to a particular functional form and its fitting algorithm does not necessarily produce a global optimum. Comparison with other types of models will measure how much of the predictable variation CART is still able to capture.

(2) Consider alternate indices. The literature offers competing measures of cognitive and motor ability. Payment formulas based on these measures may offer better estimates of cost. Furthermore, there is more data now than when the original motor and cognitive scales were developed. With this additional data we can test the relative predictive strength of the various measures and consider their practical merits.

(3) Evaluate out-of-sample performance of the models. An important element of a payment system is whether payment formulas developed from data in one year apply in future years. This is the critical measure of reliability of the derived payment system. Extrapolation can also help to determine the necessary nature and frequency of adjustments to the payment formulas as the system evolves over time.

Since the meetings, we have taken these steps and have updated our computations based on two years of additional data.

This report begins with a description of the data, followed by a discussion of modeling methods and results (which covers the above three points), then a section on obtaining new FRGs. An accompanying questionnaire seeks evaluation from the TEP on some tentative decisions we have made.

2. DATA

2.1 DATASET CONSTRUCTION

We used the merged MEDPAR/FIM data for calendar years 1996 through 1999, which contain one record for each hospital discharge. MEDPAR data describe all inpatient stays (including rehabilitation stays) paid for by Medicare. FIM data describe the functional status of patients cared for in rehabilitation facilities. Data set construction is documented for 1996 and 1997 in the project work plan (Carter, Relles, and Wynn, 2000). The same methods were applied for 1998 and 1999, and the results are discussed below. Updated documentation for data set construction is forthcoming.

Four data systems were the primary sources for the files:

- Medicare program information--the Medicare data files include discharge files recording demographic, clinical, and financial information, and hospital-level files providing facility characteristics and financial information.
- The Uniform Data System for medical rehabilitation (UDSmr). UDSmr provides functional status and demographic information for rehabilitation discharges from participating hospitals.
- The Caredata Data System (COS) for medical rehabilitation (1996-1997). Caredata also records functional status and demographic information for rehabilitation discharges from participating hospitals.
- HealthSouth Data. Caredata ceased to exist prior to our getting its 1998 and 1999 data, but we were able to obtain the data from its principal client, HealthSouth Corporation, for those years.

Our earlier MEDPAR files covered calendar years 1996 and 1997 and contained about 350,000 rehabilitation records per year. During 1996 and 1997, between 40 and 50 percent of the MEDPAR rehabilitation hospitals participated in UDSmr or Caredata. By 1999, the number of MEDPAR rehabilitation cases had grown to about 390,000, an 11 percent increase from 1996. As new hospitals joined UDSmr and HealthSouth, our

FIM sample grew even faster (by 45 percent), and now covers about 62 percent of the rehabilitation hospitals.

We used probabilistic matching methods to link records from the Medicare Provider Analysis and Review (MEDPAR) and UDSmr/Caredata/ HealthSouth (FIM) files that described the same discharge. Our merged file in 1996 matched about 55 percent of all MEDPAR rehabilitation cases. Given the steady increase in the volume of FIM cases, we now match about 70 percent of all MEDPAR cases. Our match rates remained about constant throughout. We were able to find a MEDPAR record for about 95 percent of the FIM cases where Medicare was listed as the primary payer. We matched about 90 percent of MEDPAR cases to FIM for hospitals that provided FIM data for an entire year.

2.2 DATASET CONTENTS

The merged MEDPAR/FIM data contained several variables we would need for modeling and classification. Table 2.1 identifies these variables, and indicates at which stages of the process they were used.

Table 2.1

MEDPAR/FIM Variables and Stages of Use

Purpose	Variable	Source	Description
Selection			
	AGE	MEDPAR	age
	DISSTAY	FIM	discharge stay indicator
	LOS	MEDPAR	length of stay
	IMPCD	FIM	rehabilitation impairment codes
	PROVCODE	MEDPAR	provider code
	PROVNO	MEDPAR	provider number
	TCOST	MEDPAR	total cost estimates, based on cost to charge ratios, adjusted by area wage index (*)
Clinical partitioning			
	IMPCD	FIM	impairment code
	RIC	FIM	clinical groupings resulting from impairment code mappings
Resource use			
	TCOST	MEDPAR	total cost estimates, based on cost to charge ratios, adjusted by area wage index (*)
	COGNITIVE	FIM	cognitive scores (**)
			comprehension
			expression
			social interaction
			problem solving
			memory
	MOTOR	FIM	motor scores (**)
			eating
			grooming
			bathing
			dressing--upper body
			dressing--lower body
			toileting
			bladder management
			bowel management
			bed, chair, wheelchair transfer
			toilet transfer
			tub or shower transfer
			walking or wheelchair
			stair ascending and descending
	AGE	MEDPAR	age

(*) methods described in DRU-2161-1-HCFA, Section 7.

(**) these individual components are organized into various types of indices, according to body areas and types of impairment. See Table 3.1.

The selection variables define what we think of as the typical case. We exclude transfers to hospitals and long term care settings, deaths, cases of three days or less duration, and statistical outliers. Also, the clinical partitioning and resource use variables needed to be present and in range. Selection was based on the intersection of the rules shown in Table 2.2.

Table 2.2

Rules for Selection of Modeling Cases

Variable	Selection Requirement
AGE	between 16 and 105
DISSTAY	indicates discharged to the community
LOS	more than three days, less than one year.
IMPCD, TCOST	we excluded cases with wage-adjusted log-cost more than three standard deviations from its average within RIC
PROVNO, PROVCODE	4-digit rehabilitation provider number between 3025 and 3099, or provider code = "T"
IMPCD	contained in impairment list for assignment to rehabilitation categories (see DRU-2309-HCFA, Table 3.9)
TCOST, COGNITIVE, MOTOR	greater than zero

Table 2.3 shows the amount of data we had to work with, before and after selection, by FIM source. Most of the reduction in cases is for ineligibility: deaths, interrupted stays, or transfers. The last column indicates how many cases were kept with full information. Overall, the reductions due to missing cost data and data quality (present and in-range, exclude cost outliers) are small: about 3 percent in 1996, 4 percent in 1997, 2 percent in 1998, and 3 percent in 1999. Fortunately, the additional reduction due to cost outliers is especially small--less than 0.3 percent everywhere--so we do not believe we are contaminating our results by the cost outlier exclusions.

Table 2.3

Number of Linked MEDPAR/FIM Records

Calendar Year	Source	Initial Number of Records	Rehabil-itation Facility	Present and In-Range	Eligible	Exclude Cost Outliers
1996	Total	188889	171626	166645	127595	127276
1997	Total	222682	206032	197076	149350	148966
1998	Total	246450	232691	228248	170266	169816
1999	Total	273548	257024	249941	187257	186766

Our numbers of 1996 and 1997 cases are slightly reduced from the numbers shown in the interim report, Table 3.2. The reason is that in 1996 and 1997 we were only working with the standard motor and cognitive indices, and had imputed their values from partial information if available. Here, because we needed to work with individual components, and several alternative sub-scales, we eliminated all cases that were not complete on all components. This reduced our 1996 counts by about 300 and our 1997 counts by about 200 cases.

2.3 CASE STRATIFICATION AND SAMPLE SIZES

Previous work had established 21 clinical groupings of patients according to rehabilitation impairment codes (RICs) within which we would be fitting models. Table 2.4 describes those groupings and the sample sizes available for the modeling effort according to the selection rules in Table 2.2.

Table 2.4

RIC Definitions and Sample Sizes

Rehabilitation Impairment Category	1996	1997	1998	1999
1 Stroke	33013	35387	37012	37340
2 Traumatic brain injury	1401	1653	1871	2053
3 Nontraumatic brain injury	2542	2874	3402	3758
4 Traumatic spinal cord	743	812	930	953
5 Nontraumatic spinal cord	3802	4356	5295	5837
6 Neurological	4755	5755	7832	8875
7 Hip fracture	16171	17341	18774	20627
8 Replacement of lower extremity joint	31169	37418	40931	43427
9 Other orthopedic	5310	6584	8022	9310
10 Amputation, lower extremity	4823	5437	5930	6156
11 Amputation, other	357	478	542	662
12 Osteoarthritis	2347	2860	3983	5036
13 Rheumatoid, other arthritis	1167	1527	1944	2350
14 Cardiac	4107	5677	6885	8104
15 Pulmonary	2451	3571	4340	5382
16 Pain Syndrome	1328	1890	2529	2993
17 MMT, no brain or spinal cord injury	1192	1302	1540	1679
18 MMT, with brain or spinal cord injury	160	224	221	256
19 Guillain-Barre	241	278	299	313
20 Miscellaneous	10126	13442	17423	21553
21 Burns	71	100	111	102
Total	127276	148966	169816	186766

3. MODELING METHODS AND RESULTS

We identify and discuss the three basic suggestions of the technical expert panel. Then we present a computational experiment to examine their implications, leading to selection of specific methods for determining and evaluating FRGs. The FRG selection itself is described in Section 4.

3.1 SUGGESTIONS OF THE TECHNICAL EXPERT PANEL

3.1.1 Explore Alternative Model Forms

We expect that classification and regression trees (CART) will form the final determination of the FRGs. According to the BBA Relief Act, the rehabilitation PPS system is to be based on discharges classified according to functional-related groups based on impairment, age, comorbidities, and functional capability of the patient as well as other factors deemed appropriate to improve the explanatory power of functional independence measure-function related groups. CART is the traditional method of generating FRGs (Stineman et al., 1997) and a reasonable method of determining rules to classify patients into groups that explain cost. Various algorithms have been proposed to build tree structured regression models, all of which tend to be minor variations on CART. CART is efficient at producing simple and effective rules for prediction but also has its limitations. We discuss the details of CART's strengths and limitations in the next section.

After computing an unbiased estimate of the predictive performance of a particular regression tree it is still difficult to judge how much better we might have done if we were not subject to CART's limitations. We know that R-squared ought to be between 0.0 and 1.0 with the highest values indicative of nearly perfect prediction. But when its score is potentially much lower than 1.0 we need a way to judge whether CART has performed as best as could be expected. To further investigate this we compared CART's performance with other methods.

We compared CART to ordinary linear least squares regression models, generalized additive models (GAM), and multiple adaptive

regression trees (MART). The first of these three methods is classic, the second is relatively new, and the last is the latest in prediction methodology. These models are all discussed in the statistical literature. We used the version of GAM (Hastie and Tibshirani, 1990, Generalized Additive Models, Chapman and Hall) implemented in the statistical package S-plus. MART is described in Friedman (2000), and we used software provided by the author.

To determine which models fit best we will assess each model's predictive performance on preceding and subsequent years. That is, we will fit each model (CART, linear regression, GAM, and MART) to 1997 data, for example, and use that model to predict cost for 1996, 1998, and 1999. The model that consistently predicts cost the best, in terms of the average squared difference between the actual and predicted cost, across the various years and RICs will be the gold standard.

3.1.2 Explore Predictive Ability of Other Functional Measures

The search for an ideal index set to predict cost occurred in two stages. First, we examined individual components. The main question was whether the components entered the model in the expected direction. More specifically we fit a linear regression model predicting cost from the components of the motor and cognitive scale. We checked to see which, if any, of the components had positive coefficients implying that greater functional independence increased cost. Such irregularities would flag further investigation of the data collection process for that component of the scale. We then might reconsider how or if it would be used in the index set. We also fit GAM to the components to look for non-linear effects.

Second, we experimented with the sub-scales described in Stineman, Jette, et al. (1997b). These split out the standard motor index into dimensions reflective of different body areas and types of function.

3.1.3 Examine Stability Across Years

Our previous results were based on 1996 and 1997 data, and did not give us much latitude for examining stability over time. We did verify, however, that the FRGs on the 1994 data predicted costs well in 1996 and 1997. With the addition of 1998 and 1999 data, we have the option of

fitting models within each year and seeing how well they do on other years. We also have the ability to pool multiple years worth of data for RICs that are small and hence would not otherwise have much out-of-sample predictive power.

3.2 COMPUTATIONAL DESIGN

3.2.1 Types of Models

We list below the types of models we fit and our reasons for fitting them. Included with each of the methods is a two-dimensional visualization of the surface that each model fits to data. The data come from RIC 01 (Stroke) combining 1998 and 1999 data. The darkest regions of the plots show the regions where the model predicts the lowest cost for the motor and cognitive score combination. Since such visualization is limited to two dimensions, the plot intentionally excludes age.

A. OLS--Ordinary least squares. Linear models are fit with ordinary least squares regression. In a linear model, a fixed amount of change in an independent variable, anywhere along its scale, results in the same change in the prediction of the dependent variable. For example, a change in the motor score from 20 to 21 would decrease predicted cost by the same percent as a change from 60 to 61. In other words, the coefficients of ordinary least squares report the increase in log-cost due to a unit increase in an individual component of the FIM measure.

Figure 3.1 demonstrates the linearity by the parallel contours. It also shows that the strongest effect is due to the motor score.

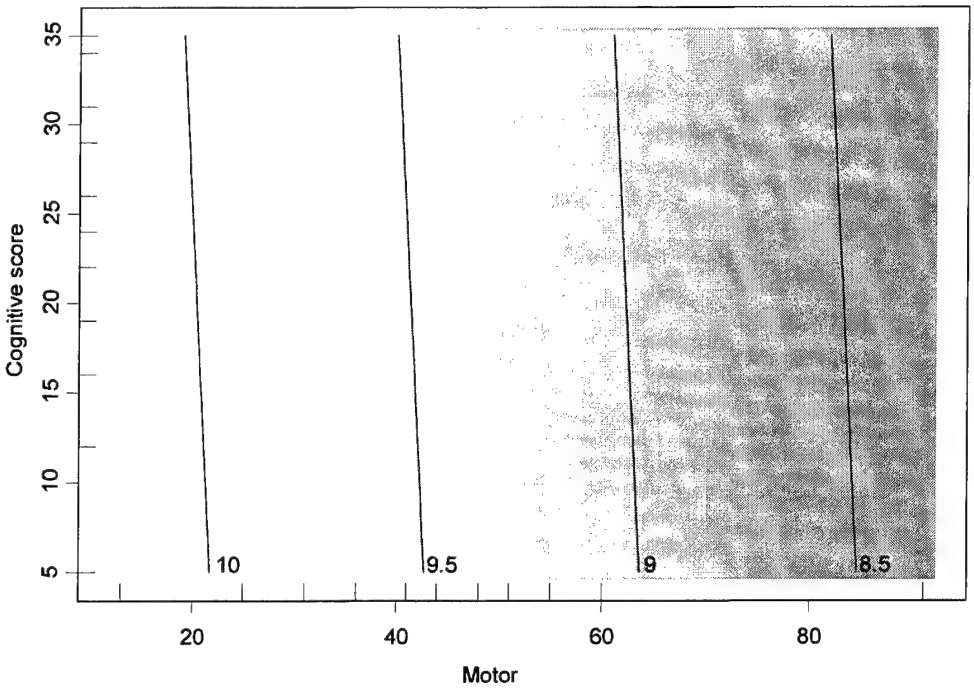

Figure 3.1—Linear Model

This compact representation allows for easy interpretation and diagnosis of the model. In particular, we looked for coefficients that indicated that increases in functional ability tended to increase cost. We flagged these components for further investigation. Besides the interpretation, OLS is computationally inexpensive and often provides an accurate approximation to the relationship between log-cost and the functional measures. It would be an appropriate gold standard if the assumption of a linear relationship between the independent variables and the dependent variable is true or approximately true.

B. GAM—Generalized additive models. GAM permits slightly more flexible relationships between the dependent and independent variables. GAM approximates the relationship as a sum of smooth (rather than linear) functions of the independent variables. This means that a change in motor score from 20 to 21 might decrease predicted cost by a different percentage than a change from 60 to 61. It does not model interactions, but only produces estimates of additive effects. Because the relationship is assumed additive, the decrease in predicted cost due to a change in motor score from 20 to 21 will be the same regardless of

the values of the other independent variables. The top two panels of Figure 3.2 compare the linear model to GAM for predicting cost from motor score. Although the two fits seem to agree closely, the GAM fit shows evidence that the effect of motor score tapers off as motor score gets smaller. The discussion of the bottom two plots is in a later section.

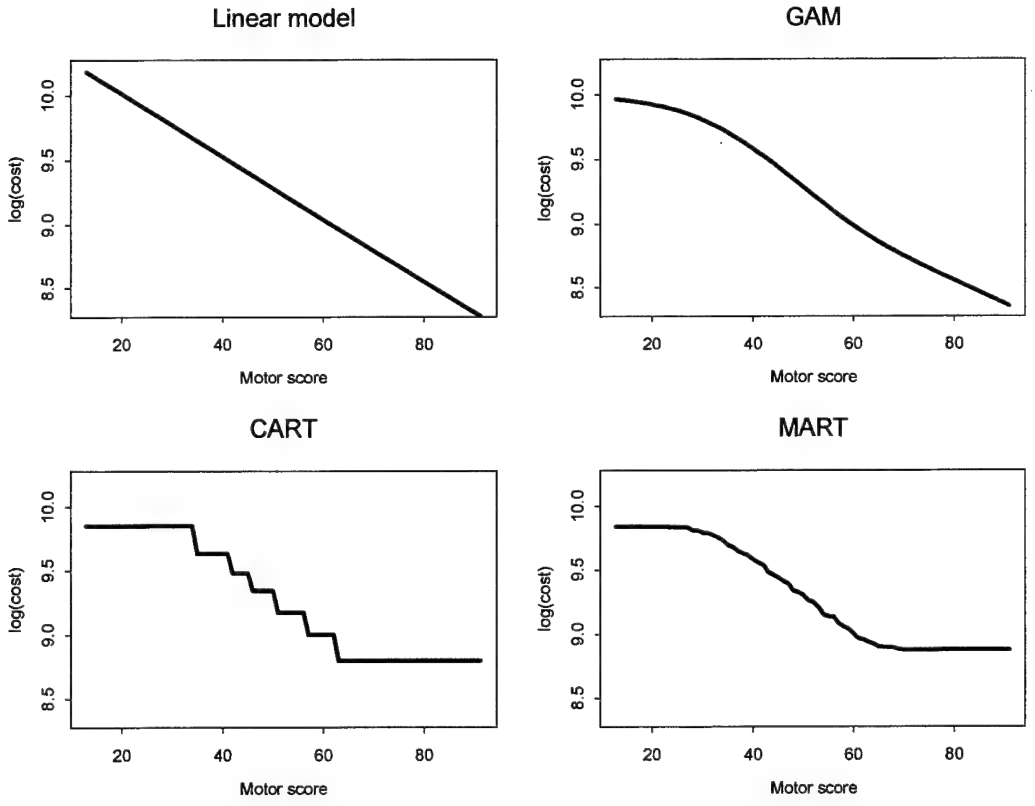

Figure 3.2—Comparison of Models for the Univariate Case

Although the additivity restriction may prevent the discovery of interaction effects in multivariate data, the benefits of additivity include easy computation and interpretation. To interpret GAM we can plot for each index the value of the index versus the contribution it makes toward the log-cost estimate. We can then visually look for irregularities, saturation effects, and threshold effects. For example, we may learn that patients with motor scores exceeding a particular value have roughly constant cost, an example of a saturation effect. GAM does use more degrees of freedom than OLS but conserves them by

imposing the additive constraint and restricting the additive components to be very smooth, spending roughly four degrees of freedom per predictor. GAM will also work well in small RICs.

Figure 3.3 shows the shape of the GAM fit. Clearly, GAM picks up curvature that the linear model cannot. It is still apparent that the motor score is the most influential. However, GAM also seems to pick up that at extreme values on the cognitive scale the cost is slightly lower than for cognitive scores in the middle of the range.

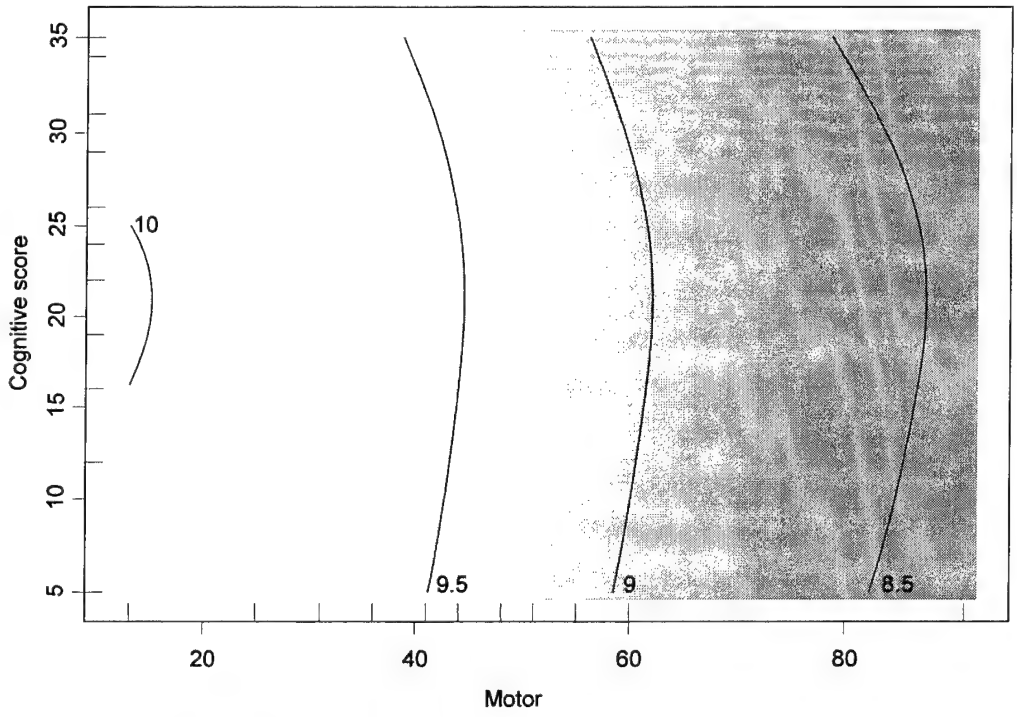

Figure 3.3—Generalized Additive Model

The cost of the additional flexibility is greater model complexity and variability. However, that same flexibility that makes GAM more complex also can make its predictions more accurate than the linear model when the relationship between the dependent and independent variables is non-linear.

C. MART—Multiple adaptive regression trees. MART is a state-of-the-art statistical method. MART is the most flexible and most complex of the models under consideration as a gold standard. Like GAM, it is nonparametric with the ability to find non-linear relationships.

However, it is also able to find interaction effects in the predictor variables.

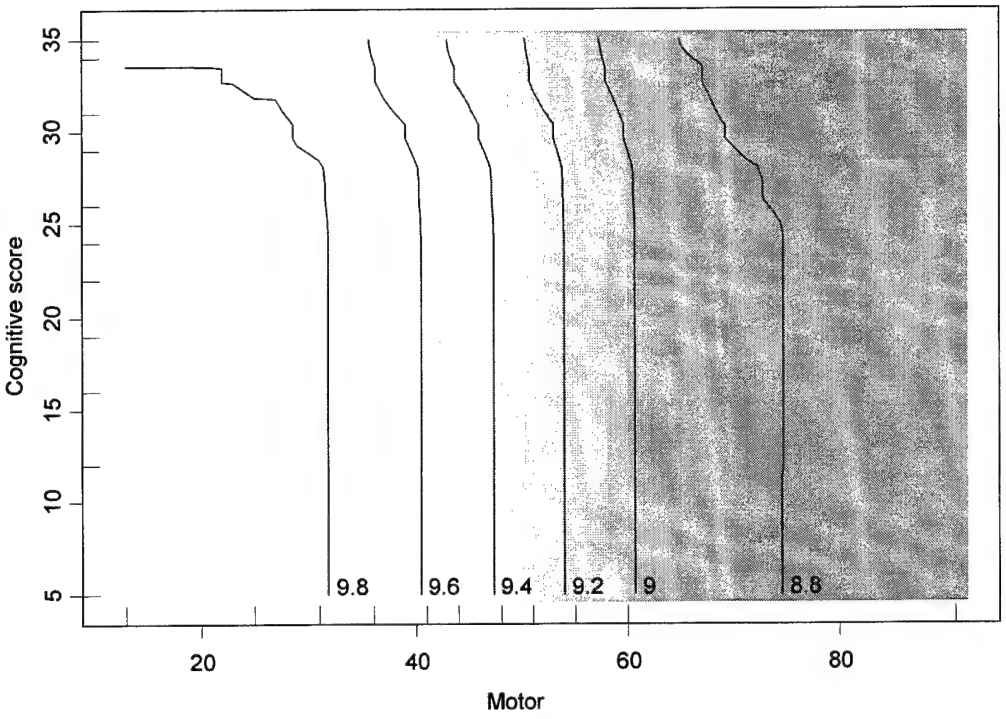

Figure 3.4—Multiple Adaptive Regression Trees

The MART prediction is the sum of predictions from many simple CART models. The algorithm constructs the CART models sequentially in such a way that each additional CART model reduces prediction error. Since each CART model fits an interaction effect, the sum of many of them (100s to 1000s) results in a prediction model that permits complex, non-linear relationships between the dependent and independent variables. We can control the depth of interaction effects MART tries to capture by controlling the depth of the individual CART models.

If cost varies in a non-additive way across motor score and cognitive scores, then MART might be able to capture this information and provide predictions that are more accurate than GAM. Figure 3.4 shows the shape of the MART fit. Like GAM, MART determines that in the high cognitive values have lower costs than the lower cognitive scores at a fixed motor score. Furthermore, MART shows that costs decrease much faster at the high cognitive scores for very low motor scores.

This is a feature that the functional form of GAM cannot detect. When such effects are strong then MART would likely outperform GAM. This makes it a good candidate for the gold standard.

Like GAM, the additional complexity complicates interpretation. It is difficult to interpret and it is difficult to quantify the number of degrees of freedom that it spends. However, some measures of variable influence and visualization tools are available for evaluating the predictor's rationale. It is not clear if it will always work well for very small RICs, but results show that it has been competitive with GAM.

D. CART—Classification and regression trees. CART is a well-known technique for building classification models (Breiman et al., 1984). CART requires a dependent variable (here, log-cost), and it seeks to develop predictors of the dependent variable through a series of binary splits from a candidate set of independent variables (here, age, FIM motor score, and FIM cognitive score). CART partitions the data into two groups according to the independent variables. Such a partition might separate patients with motor score exceeding 50 from those with motor score less than 50. CART chooses the variable on which to split the data and the value of the variable at which to split so that the new partitions are more homogeneous in terms of log-cost. The partition minimizes the squared prediction error. CART then recursively splits each partition until it satisfies some stopping criteria.

Figure 3.5 shows how CART partitions in the data example. Figure 3.6 shows the dendrogram (tree) version of the plot. The findings are not unlike those of the previous analysis. We can still see that motor is the primary effect although at high motor scores cognitive ability can be influential. The linear model, or GAM, cannot detect such interaction effects. However, the boundaries are abrupt and discontinuous. This is particularly noticeable in Figure 3.2. There, the CART fit seems to be trying hard to fit a fairly smooth function yet is necessarily jagged. This kind of bias is likely to detract from the fit relative to GAM or MART.

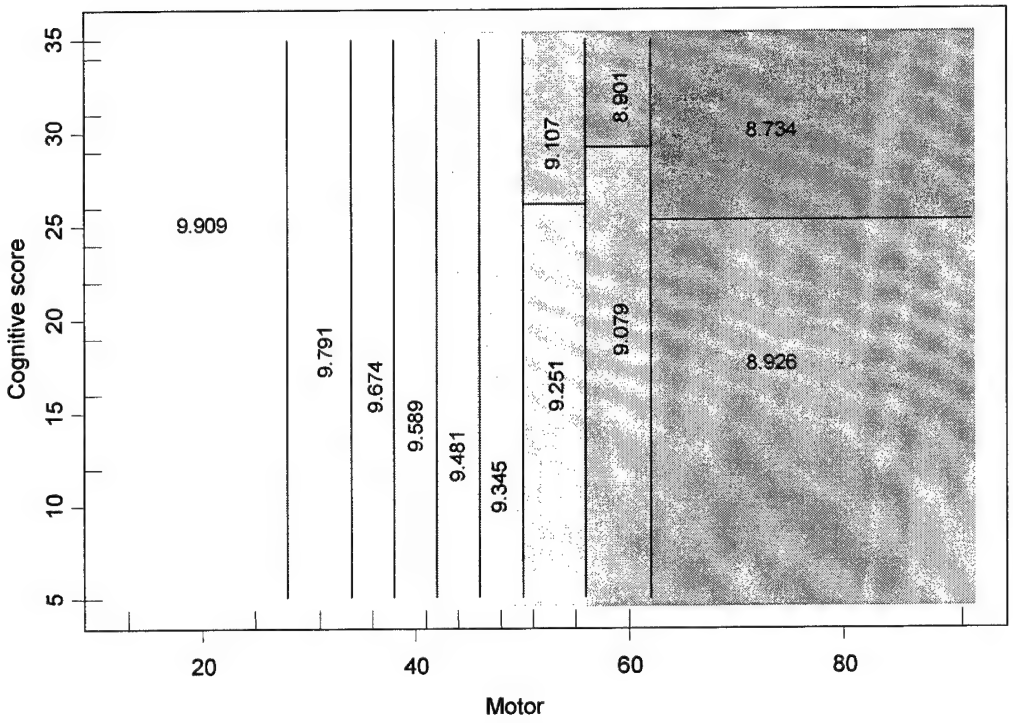

Figure 3.5—Classification and Regression Tree

By recursively partitioning the data, CART essentially fits interaction terms and thus can miss some main effects. CART has the pleasing theoretical property that as the sample size grows the prediction rule converges to the one that minimizes the expected prediction error. However, even in large finite samples CART can fail to fit curvature well (underfit) or can infer curvature where none exists (overfit). CART is a high-variance regression method, meaning that small fluctuations in the data set can produce very different tree structures and prediction rules. An early split will influence the shape of the tree and produce results that may be nonsensical. In practical use with large data sets, CART can produce a tree with many partitions causing difficulty in interpretation and evaluation of the inferred rules.

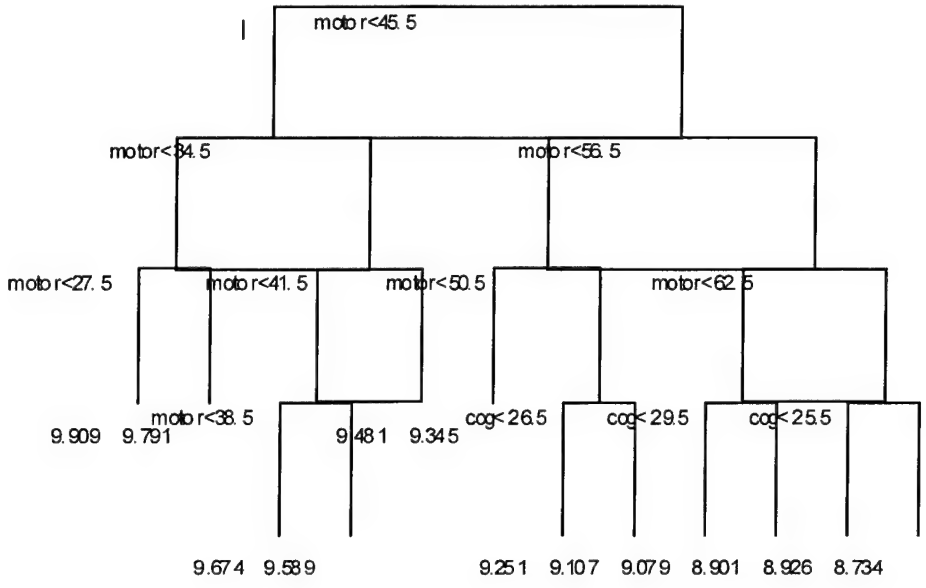

Figure 3.6—Dendrogram of the CART Model

With knowledge of these limitations, CART can still be a useful and powerful tool. The CART model offers the advantage of producing groups determined by ranges of the independent variables. It becomes easy to classify a new patient by comparing the values of the patient's set of independent variables with the ranges that define each of the CART determined groups. Our use of CART focused on three stopping criteria, all of which attempt to estimate the optimal number of partitions to generate from the data.

D.1. XVAL--CART, standard cross-validation to minimize MSE. CART's performance can be very sensitive to the number of partitions it produces. Too few partitions fail to separate patient groups with very different associated costs. Too many partitions cause the cost estimates to be unreliable as well as causing difficulty in practical implementation of the payment formula. Somewhere in between is the optimal number of partitions.

We used ten-fold cross validation, the most widely used method, to estimate the number of partitions. This method splits the data into ten groups containing equal numbers of patients. For each of the splits we construct a CART model on the other 90 percent of the observations and evaluate the performance of various tree sizes on the current validation split. We then average the performance over the ten validation runs by tree size. We select the tree size with the lowest cross-validated mean squared error to be the best tree size. We fit a final single CART model to the entire data set, stopping when the tree size reaches the ten-fold cross-validation choice.

D.2 1SD--CART, stop when within one standard deviation of minimum MSE. Since we are working with fairly large data sets, it turns out that the ten-fold cross-validation method can produce models with far too many splits. We needed to introduce "practical" considerations into the stopping criteria. Breiman et al. (1984, 78-80) recommended a more aggressive stopping rule to fix the number of partitions that corrects this situation. They suggest placing confidence bands around the cross-validated estimate of prediction error by tree size. Then choose the first node where prediction error is within one standard error of the minimum. This reduces the number of partitions, reduces the probability of overfitting, and could cause some more heterogeneous groups (in terms of log-cost) to be combined.

D.3 INT--CART, interim report numbers of nodes. This algorithm simply fixes the number of nodes to be those used in the project's Interim Report (DRU-2309-HCFA, July 2000, Table 3.12). This stopping criterion is based on 1997 data and is useful for comparison with the high variance methods that must estimate the tree size in combination with the partitions and payment levels for each partition.

3.2.2 Alternative Functional Impairment Indices

Table 3.1 shows the FIM items that we considered as independent variables. From these we assembled various indices by summing item responses to determine whether predictive strength varied across items or groups of items.

Table 3.1

The Candidate Indices

<table>
<tr><th>Items</th><th>M13C5</th><th>M12C5</th><th>M12C4</th><th>StJe3</th><th>StJe5</th></tr>
<tr><td>transfer to tub/shower</td><td rowspan="12">standard motor</td><td>X</td><td>X</td><td>X</td><td>X</td></tr>
<tr><td>transfer to bed/chair</td><td rowspan="12">motor excluding trftub</td><td rowspan="12">motor excluding trftub</td><td rowspan="4">mobility excluding trftub</td><td rowspan="2">transfer excluding trftub</td></tr>
<tr><td>transfer to toilet</td></tr>
<tr><td>Walking/ wheelchair</td><td rowspan="2">locomotion</td></tr>
<tr><td>stairs</td></tr>
<tr><td>bladder</td><td rowspan="6">ADLS</td><td>sphincter</td></tr>
<tr><td>bowel</td><td rowspan="6">self care</td></tr>
<tr><td>eating</td></tr>
<tr><td>grooming</td></tr>
<tr><td>bathing</td></tr>
<tr><td>dress upper</td></tr>
<tr><td>dress lower</td></tr>
<tr><td>toilet</td></tr>
<tr><td>comprehension</td><td rowspan="6">standard cognitive</td><td rowspan="6">standard cognitive</td><td>X</td><td rowspan="6">standard cognitive</td><td rowspan="6">standard cognitive</td></tr>
<tr><td>expression</td><td rowspan="5">cognitive excluding compreh</td></tr>
<tr><td>social interaction</td></tr>
<tr><td>problem solving</td></tr>
<tr><td>memory</td></tr>
</table>

Note: transfer to tub has been a traditional component of all these mobility indices. However, for reasons developed in Section 3.3.1, we take transfer to tub out of the relevant indices when the time comes to use them.

A. Components. The motor FIM scale contains 13 items and the cognitive FIM contains five. The Component index set allows each of the individual items to contribute to the model as independent variables. We wanted to try all of the 18 responses to see what additional information they might provide. We hypothesized that there would not be enough information in the data to fit interactions among these. Indeed, when we initially fit CART models on all indices, we were getting very poor fits. We thought, however, that additive models might give useful information on their relative contributions, and we attempted to apply both OLS and GAM to these.

One other piece of information that we wanted to examine was the indicator of mode of locomotion: whether wheelchair or walking. We experimented with a wheelchair dummy variable and an interaction between the wheelchair indicator and the walk/wheelchair response in our OLS and GAM fits.

B. M13C5—Standard two scales (motor + cognitive). These were the standard indices that we had used for the interim report. Only two terms are in this index, the sum of the 13 responses to the motor FIM items and the sum of the five responses to the cognitive FIM items.

C. M12C5—Standard, but exclude transfer to tub/shower in motor score. This index set arose during the course of our investigation of individual items. As discussed further in Section 3.3.1, we found that patients with greater functional independence in transfer to tub/shower tended to cost more. Thus it is reasonable to believe that eliminating this item from the motor score may improve the prediction of cost.

Only two terms are in this index, the sum of 12 responses to the motor FIM (transfer to tub excluded) and the sum of the five responses to the cognitive FIM.

D. M12C4—Standard, exclude tub transfer in motor score and comprehension in cognitive score. This index set also arose during the course of our investigation of individual scores (see Section 3.3.1). In the cognitive FIM, increased functional independence on the comprehension component tended to increase cost. Consequently, we explore eliminating the comprehension item from the cognitive scale, although eliminating this item may be undesirable for reasons discussed further in Section 3.3.3.

Only two terms are in this index, the sum of 12 of the responses to the motor FIM (transfer to tub excluded) and the sum of four of the responses to the cognitive FIM (comprehension excluded).

E. StJe3—Stineman and Jette, Activities of daily living (ADLS, mobility, standard cognitive). This and the following index set were proposed by Stineman et al. (1997a) as sub-scales of the cognitive and motor score that might relate to specific impairments. This index was found to describe dimensions of function within the large stroke RIC. Our question is whether there is additional information in these indices

that could help to predict cost and to improve on the classification system, either in some or in all RICs. We set out to determine their potential contribution to cost prediction. Initially, we developed these indices as defined by Stineman. However, once we determined that we preferred the motor index without tub transfer (item C, above), we defined its corresponding component mobility/transfer to exclude tub transfer as well.

This index set has three components, the sum of four of the mobility components of the motor FIM (transfer to tub excluded), the sum of the eight daily living components of the motor FIM, and the sum of all five of the components of the cognitive FIM.

F. StJe5—Stineman and Jette, four motor scores (self-care, sphincter, transfer, locomotion) plus standard cognitive. This index set is a further decomposition of the previous set. It breaks down the ADLS index into self-care and sphincter and decomposes the mobility index into transfer and locomotion subindices. They were found by Stineman and Jette to be dimensions of function in RICs 6 through 14, 17, 19, and 20.

This index set has five components, the sum of two of the transfer components of the motor FIM (transfer to tub excluded), the sum of the two locomotion components of the motor FIM, the sum of the two sphincter control components of the motor FIM, the sum of the self-care components of the motor FIM, and the sum of all five of the components of the cognitive FIM.

3.2.3 Fitting and Evaluation Periods

To validate the various estimators of the relationship between the indices and log-cost we evaluate each method in terms of out-of-sample predictive performance. The most important fits, of course, are the ones based on the most recent data, for they will determine the payment system. We can get an idea of how well they will perform by seeing how earlier years' fits perform on following years' data. We initially tried fitting separate models for each year and seeing how well they performed on all other years. This would yield 12 out-of-sample fits/evaluations. We later improved that by observing that some RICs

(e.g., 04, 11, 18, 19, 21) were quite small, and it might be advantageous to pool their data. This led to experimenting with fitting periods 1996-97 and 1998-99. Thus, the full set of fits and predictions is described by the appearance of "x" in Table 3.2.

Table 3.2

Combination of Fitting and Evaluation Periods Examined

	Evaluation Period			
Fitting Period	**1996**	**1997**	**1998**	**1999**
1996	.	x	x	x
1997	x	.	x	x
1998	x	x	.	x
1999	x	x	x	.
1996-97	.	.	x	x
1998-99	x	x	.	.

3.3 RESULTS

3.3.1 Item Level Analysis

We ran OLS with individual sub-scales (eating, walking, etc.). We wanted to know whether the individual items appeared to influence costs in the expected direction: higher FIM scores should mean lower costs. OLS with log-cost would be the easiest method to interpret. If the estimated coefficients were positive, a variable's effect would be inconsistent with clinical expectations.

Randomness alone would produce numerous positive regression coefficients: there are 18 individual components, 21 RICs, coefficients can be measured imprecisely, and so we expect a number of small t-statistics that could be on either side of zero. However, we have good power to detect if a given effect is consistently positive. If the pattern of positive signs persists for all four years of data and for several RICs within each year, we would have some confidence that found an anomalous item.

Table 3.3 shows for the OLS regressions how many RICs had a positive sign within each of the data sets across the 21 RICs, and how many these coefficients had t-statistics greater than 1.0. The unmistakable patterns are that both tub transfers and comprehension

often have the wrong sign in OLS regressions--costs were higher when the functional independence measure was higher.

Table 3.3

Component Regressions:
Occurrences of Positive Regression Coefficients in 21 RICs

	Positive OLS Coefficients						OLS t-statistic>=1.0					
Variable	**1996**	**1997**	**1998**	**1999**	**96-97**	**98-99**	**1996**	**1997**	**1998**	**1999**	**96-97**	**98-99**
comprehension	19	15	19	19	20	20	13	11	15	15	15	16
expression	7	4	6	7	4	4	3	1	2	1	1	3
social interaction	14	10	7	9	11	7	6	3	1	3	5	3
problem solving	3	4	4	7	3	4	0	1	2	2	0	2
memory	5	5	4	2	4	1	3	2	0	1	1	0
eating	1	0	1	0	0	1	1	0	0	0	0	0
grooming	9	11	13	12	8	12	5	7	9	7	6	10
bathing	5	2	2	1	2	2	0	1	1	0	0	0
dress upper body	13	10	12	14	11	15	3	3	7	10	6	10
dress lower body	0	3	1	1	1	0	0	1	0	0	1	0
toileting	0	0	1	1	0	1	0	0	0	0	0	0
bladder	1	1	1	3	1	0	0	0	0	0	0	0
bowel	11	6	14	12	7	13	4	5	3	5	6	6
transfer to bed	2	2	0	0	1	0	0	1	0	0	1	0
transfer to toilet	2	1	0	2	0	0	1	0	0	0	0	0
walking	1	0	0	0	0	0	0	0	0	0	0	0
stairs	5	3	5	4	2	4	2	2	2	2	2	1
transfer to tub	17	19	21	18	20	20	14	13	18	16	13	18

We believe that the perverse effect of transfer to tub is due to the fact that the response depends on the situation being scored—either tub or shower and with or without assistive devices. The UDSmr question and answer manual says: "It may be that a subject's score goes down as he/she no longer requires the use of some assistive device" (p. 31). It is likely that patients who do not use a device at admission score worse than patients who do—so the FIM item provides only a situational measure of independence rather than an absolute measure.

We have no similar rationale for the comprehension results. It may be that this finding reflects only that many hospitals do more for patients that understand what is happening. However, eliminating this item raises issues related to incentives and fairness. If we take the comprehension item out of the index, the system will provide no extra incentives to treat patients with lowered comprehension. If some hospitals do spend extra to treat such patients, they will not be compensated for such extra resources. On the other hand, if this really represents the current pattern of best care, our system should reflect it.

To confirm that OLS was not overlooking important non-linear effects, we also looked at plots of the marginal contributions of each component, as estimated by GAM. Although log-cost did not always smoothly decline with the seven-point scale, there were only two items where the relationship was perverse. These were the same ones that showed up in the linear models: comprehension and transfer to tub.

One other piece of information that we wanted to examine was the mode used in the walk/wheelchair item. We experimented with a wheelchair dummy variable and an interaction between the wheelchair dummy and the FIM walk/wheelchair response in our OLS and GAM fits. We found that, in most RICs, wheelchair patients cost more than expected given their functional scores, and locomotion score is less important when in a wheelchair than when walking. The net effect of wheelchair alone was quite small. However in the two RICs that have the most wheelchair people (RICs 4 and 10), neither wheelchair functional status nor the wheelchair indicator is significant. Thus, adding these variables will not result in a substantial improvement in prediction of cost.

3.3.2 Selecting a Gold Standard Model

Having a gold standard model does two things. First, it helps us understand how well CART is doing--it gives us a measure of attainable residual standard deviation, to compare to the residual standard deviation we get from CART. Second, it will enable us in a simulation exercise to assess the prediction bias for various combinations of

demographic and hospital factors. The latter simulations will be performed in the project's final report.

MART and GAM are the candidates for gold standard status. We have theoretical reasons to prefer MART. It is extremely flexible, and it detects interactions. However, its prediction formula is rather unwieldy. Also, some RIC sample sizes are small, and it may be that without forcing some structure, one effectively fits too many parameters and gets a model that does not extrapolate very well. On the other hand, GAM uses fewer degrees of freedom, and produces a curve to describe each input variable's effects, so it is a little easier to decide whether the GAM fits make clinical sense. Without a clear a priori winner, we decided to perform our computations on both GAM and MART.

Knowing that CART would not produce reasonable models with just component scores, we chose not to work further with the components at this point. We fit all combinations of models and remaining indices (six types of models, five types of indices, six fitting periods). We looked at out-of-sample root mean squared prediction error (RMSE) as a measure of quality of fits. Aggregate RMSEs across RICs are provided in Table 3.4.

Table 3.4

Root Mean Squared Errors Among Candidate Gold Standard Models

Fit Yr	Eval Yr	Const	M13C5-GAM	M13C5-MART	M12C5-GAM	M12C5-MART	M12C4-GAM	M12C4-MART	StJe3-GAM	StJe3-MART	StJe5-GAM	StJe5-MART
96	97	.541	.475	.474	.474	.473	.474	.473	.471	.470	.467	.467
	98	.545	.480	.480	.479	.479	.479	.479	.475	.475	.473	.473
	99	.546	.483	.484	.482	.482	.482	.482	.479	.479	.476	.476
97	96	.536	.469	.468	.468	.467	.468	.467	.465	.465	.462	.462
	98	.545	.480	.479	.479	.478	.479	.478	.475	.475	.472	.471
	99	.546	.483	.483	.482	.482	.482	.481	.478	.478	.476	.475
98	96	.536	.469	.469	.468	.468	.468	.468	.465	.465	.463	.463
	97	.541	.475	.474	.474	.473	.474	.473	.471	.470	.468	.467
	99	.546	.482	.482	.481	.480	.481	.481	.477	.477	.475	.474
99	96	.536	.470	.470	.469	.469	.469	.468	.466	.466	.463	.463
	97	.541	.475	.475	.474	.474	.474	.473	.471	.471	.468	.467
	98	.545	.480	.479	.479	.478	.478	.478	.475	.474	.472	.471
96-97	98	.545	.480	.479	.479	.478	.479	.478	.475	.474	.472	.471
	99	.546	.483	.483	.482	.481	.481	.481	.478	.478	.476	.475
98-99	96	.536	.469	.469	.468	.468	.468	.468	.465	.465	.463	.462
	97	.541	.475	.474	.474	.473	.474	.473	.471	.470	.467	.466

The RIC constant column fits means to RICs (i.e., one FRG per RIC). Thus it is the within RIC variance of the log of the cost of cases, using a case-weighted average across RICs. It measures the amount of variance that might be explained by defining FRGs within each RIC. Then, for each index set, we show RMSEs for GAM and MART.

There are several interesting things to observe. First, MART sometimes does a tiny bit better than GAM, but they do about equally well. Second, the index without transfer to tub (M12C5) does slightly better than the index with it (M13C5) in at least one model in all combinations of fitting and prediction years. Comparing M12C5 to the similar model without the comprehension item (M12C4), we find dropping comprehension improves prediction in only six of the 16 predictions that we evaluated. Third, both GAM and MART seem to be able to make use of additional index information. RMSE goes down as the number of indices goes up, and the RMSE is lowest for the most disaggregated set of indices StJe5. Finally, the RMSEs are all large, even for StJe5. About 15 percent of the standard deviation, or 25 percent of the variance, is

explainable. But we cannot do better than that by creating FRGs. Case level costs are inherently unpredictable.[1]

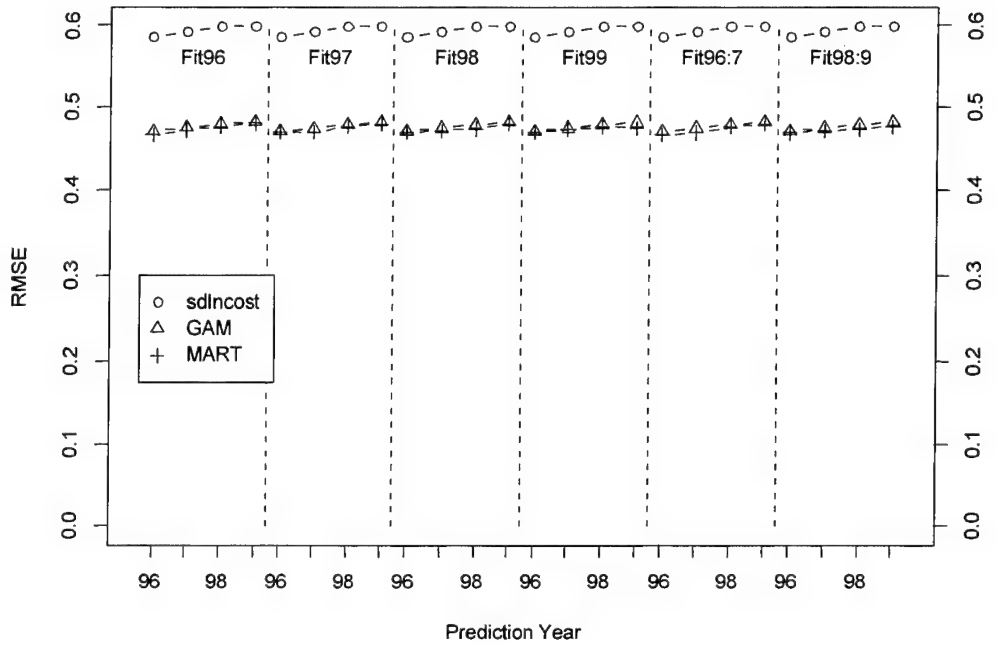

Figure 3.7—RMSEs by Fit and Prediction Years: RIC=01, Index+StJe5

[1] The payment system, of course, also exploits the variance across RICs in cost. About 34 percent of the total variance in the wage adjusted cost of cases discharged to the community is predicted by the FRG system and 37 percent by our gold standard models.

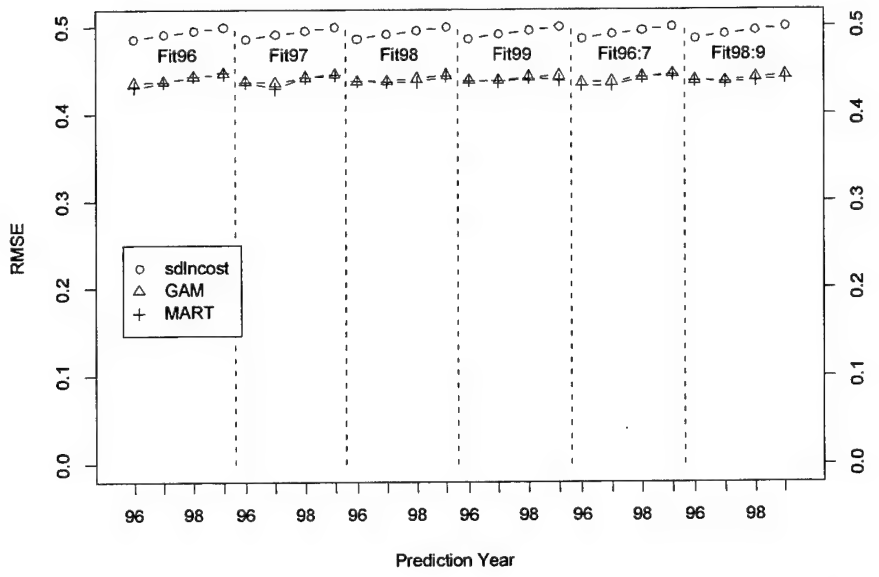

Figure 3.8—RMSEs, by Fit and Prediction Years: RIC=07, Index=StJe5

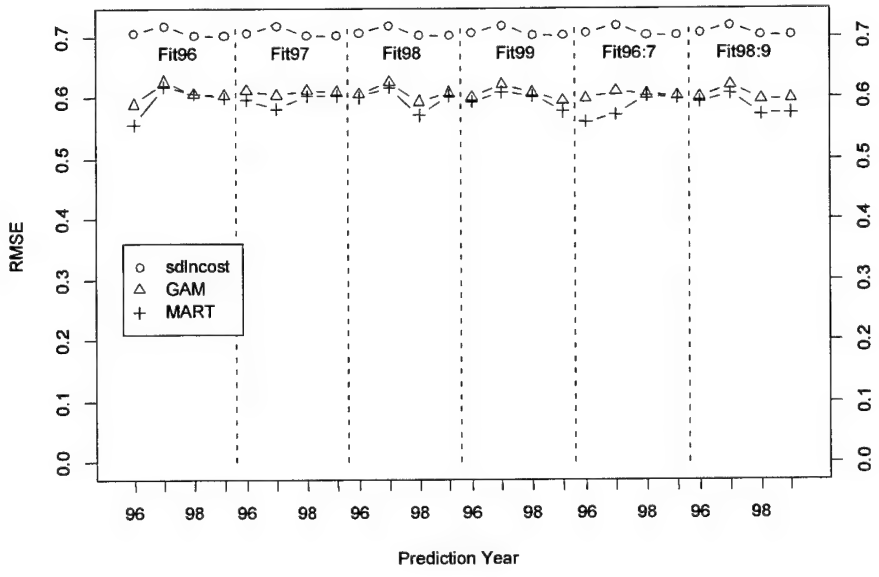

Figure 3.9—RMSEs, by Fit and Prediction Years, RIC=04, Index+StJe5

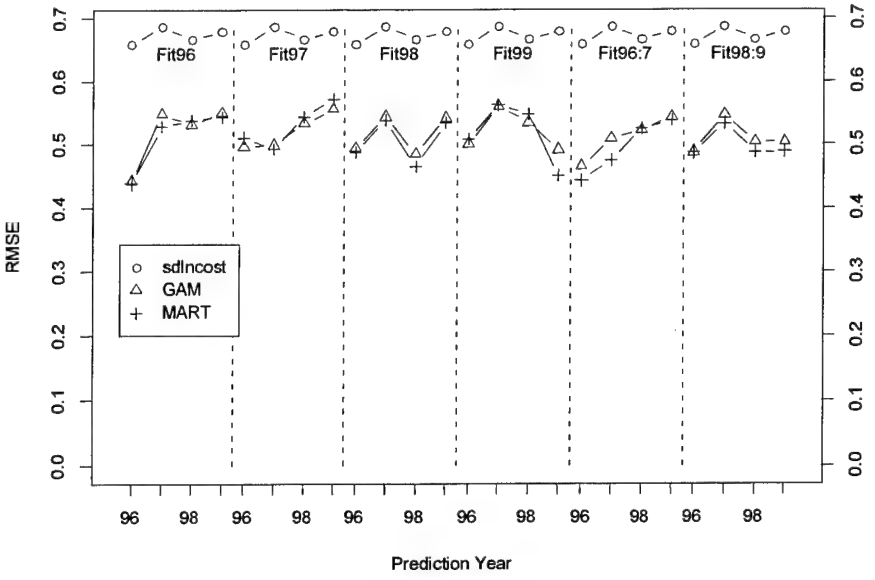

Figure 3.10—RMSEs, by Fit and Prediction Years: RIC=18, Index=StJe5

We also looked at reductions in standard deviation within each RIC. Percentage reductions varied from about 20 percent for stroke (RIC 01) to about 10 percent for the three orthopedic RICs (07, 08, 09). These orthopedic RICs are substantially more homogeneous in cost than other RICs, so that despite the fact that we predict a smaller fraction of the variance in these RICs, they have RMSEs that are among the lowest of all other RICs. Figures 3.7 and 3.8 show what was typical of most RICs: MART and GAM perform about equally well, but MART does a little better across prediction periods. Performance in the smaller RICs was similar (see Figures 3.9 and 3.10), although sometimes GAM did a little better. But in no case did GAM seem to dominate MART across prediction periods.

In summary, we saw that MART seemed to be a little better than GAM for many RICs, validating the observations we made above at the aggregate level. From the standpoint of determining the percent of explainable standard deviation, we decided to use MART with index set StJe5 as the gold standard. Prior to reviewing exactly which models to use within each RIC for our final report simulations, we assume that this model will provide a good estimate of the percent of explainable standard deviation attained by our CART models.

3.3.3 Evaluation of CART Models

We have shown above that we achieve the best prediction using MART and the StJe5 index set. But this does not lead to a simple payment system, and may meet the definition of patient groups found in the law. We would need a complex computer program to evaluate the formula. This is not compatible with the design criteria of the payment system. CART is therefore employed to produce simple, understandable patient groups.

We introduce the alternative CART models we considered by reviewing the results for the index set used in our interim report--M12C5. Table 3.5 shows RMSEs for the M12C5 model. The further to the right, the more FRGs in the CART model. The RIC constant column fits means to RICs (i.e., one FRG per RIC) and repeats the data from Table 3.4. The interim report used the one standard deviation rule with some adjustments; it performs similar to the 1SD rule applied here. The XVAL column shows how well CART does with its standard cross-validation stopping rule, which tends to produce more than twice the terminal nodes as 1SD.

Table 3.5

Performance of Alternative CART Models: Index = M12C5

		Within-RIC Standard Deviations					Percent of SD Explained (*)			Number of Nodes		
Fit Yr	Eval Yr	Const	INT	1SD	XVAL	StJe5-MART	INT	1SD	XVAL	INT	1SD	XVAL
96	97	.541	.480	.480	.477	.467	82.4	82.4	86.5	104	101	239
	98	.545	.486	.486	.483	.473	81.9	81.9	86.1			
	99	.546	.489	.489	.486	.476	81.4	81.4	85.7			
97	96	.536	.474	.473	.471	.462	83.8	85.1	87.8	104	101	239
	98	.545	.485	.485	.482	.471	81.1	81.1	85.1			
	99	.546	.488	.488	.485	.475	81.7	81.7	85.9			
98	96	.536	.474	.474	.472	.463	84.9	84.9	87.7	104	123	303
	97	.541	.479	.479	.477	.467	83.8	83.8	86.5			
	99	.546	.487	.486	.484	.474	81.9	83.3	86.1			
99	96	.536	.475	.474	.472	.463	83.6	84.9	87.7	106	119	288
	97	.541	.480	.479	.477	.467	82.4	83.8	86.5			
	98	.545	.484	.483	.481	.471	82.4	83.8	86.5			
96-97	98	.545	.485	.483	.481	.471	81.1	83.8	86.5	104	137	351
	99	.546	.488	.486	.484	.475	81.7	84.5	87.3			
98-99	96	.536	.474	.472	.471	.462	83.8	86.5	87.8	108	176	478
	97	.541	.479	.477	.476	.466	82.7	85.3	86.7			

(*) The percent of standard deviation explained by the model, where 0 equals the constant term model, and 100.0 equals the gold standard.

The main observation is that the CART models traverse a substantial fraction of the distance between the constant model and the gold standard. The CART FRGs explain more than 80 percent of the explainable standard deviation. The 1SD model, which has less than half the nodes of XVAL, explains almost as much as XVAL.

Table 3.6 compares the performance of the alternative CART models relative to the gold standard for all of the indices we considered. It shows once again that M12C5 outperforms M13C5. Considering the results for the INT models, which force an equal number of nodes, we notice that M12C5 does better per node than either StJe3 or StJe5. M12C4 performs slightly better than M12C5, but also worse in two of the prediction model pairs.

Table 3.6

Performance of Alternative CART Models: Percent of SD Explained

		Index=M12C5			Index=M12C4			Index=M13C5			Index=StJe3			Index=StJe5		
Fit Yr	Eval Yr	INT	1SD	XVAL	INT	1SD	XVAL	INT	1SD	XVAL	INT	1SD	XVAL	INT	1SD	XVAL
96	97	82.4	82.4	86.5	83.8	82.4	86.5	81.1	81.1	85.1	79.7	82.4	86.5	81.1	85.1	87.8
	98	81.9	81.9	86.1	83.3	81.9	86.1	81.9	80.6	84.7	80.6	84.7	88.9	81.9	84.7	87.5
	99	81.4	81.4	85.7	82.9	81.4	85.7	81.4	80.0	84.3	80.0	82.9	87.1	80.0	82.9	87.1
97	96	83.8	85.1	87.8	83.8	83.8	87.8	81.1	82.4	86.5	81.1	85.1	87.8	81.1	85.1	89.2
	98	81.1	81.1	85.1	81.1	81.1	85.1	79.7	79.7	83.8	79.7	83.8	86.5	79.7	83.8	87.8
	99	81.7	81.7	85.9	81.7	81.7	85.9	80.3	80.3	84.5	80.3	83.1	87.3	78.9	83.1	88.7
98	96	84.9	84.9	87.7	84.9	86.3	89.0	82.2	83.6	86.3	80.8	84.9	89.0	82.2	86.3	89.0
	97	83.8	83.8	86.5	82.4	83.8	87.8	81.1	82.4	85.1	79.7	83.8	87.8	79.7	86.5	89.2
	99	81.9	83.3	86.1	81.9	83.3	86.1	79.2	80.6	84.7	79.2	84.7	87.5	79.2	86.1	88.9
99	96	83.6	84.9	87.7	84.9	86.3	89.0	82.2	83.6	86.3	80.8	84.9	87.7	79.5	84.9	89.0
	97	82.4	83.8	86.5	83.8	83.8	87.8	81.1	82.4	85.1	79.7	83.8	87.8	79.7	86.5	89.2
	98	82.4	83.8	86.5	82.4	83.8	86.5	79.7	81.1	85.1	79.7	85.1	87.8	81.1	86.5	90.5
96-97	98	81.1	83.8	86.5	81.1	83.8	86.5	79.7	82.4	85.1	79.7	85.1	89.2	81.1	87.8	90.5
	99	81.7	84.5	87.3	81.7	84.5	87.3	80.3	83.1	85.9	80.3	85.9	90.1	80.3	87.3	90.1
98-99	96	83.8	86.5	87.8	83.8	86.5	89.2	82.4	85.1	87.8	79.7	86.5	89.2	81.1	89.2	90.5
	97	82.7	85.3	86.7	82.7	85.3	88.0	80.0	84.0	85.3	78.7	86.7	89.3	80.0	88.0	90.7

Having settled upon CART for the payment formula we are left to decide between the candidate indices. The criteria are quality of fit and parsimony. In the interim report we used 1SD-CART and 92 nodes. If we use 1SD-CART with the multiple indices in StJe3 and StJe5, CART produces far too many nodes. Raw 1SD-CART numbers of nodes in 1999 have 186 nodes for StJe3 and 201 nodes for StJe5 with almost no improvement in RMSE. If we fit with interim report number of nodes, we get RMSEs for StJe3 and StJe5 that are larger.

In CART, the index with transfer to tub (M13C5) does noticeably worse than the index without this item in many years and never does substantially better. This is similar to our findings with GAM and MART. Because we believe this item does not measure and absolute level of function, we propose to recommend to HCFA that this item not be used in creating FRGs. So, the choice seems to be between M12C5 and M12C4. The argument for going to M12C4 is that comprehension seems to work opposite to the standard cognitive scale in which it is embedded. After fixing a stopping rule, dropping comprehension from the index produces a slightly better prediction in some years. Further, the reduced

cognitive scale does not increase the frequency with which FRGs are defined by cognitive function. Eliminating the comprehension item raises issues related to incentives and fairness. If we take the comprehension item out of the index, the system will provide no extra incentives to treat patients with lowered comprehension. If some hospitals do spend extra to treat such patients, they will not be compensated for such extra resources. On the other hand, if this really represents the current pattern of best care, our system should reflect it. The improvements in predicting cost are so slight that it seems to us that the decision should be based on clinical judgment about what should be paid for. Consequently, we are requesting your opinion on the right policy choice concerning this item.

3.3.4 Cost Patterns

We wanted to understand the marginal contribution of motor and cognitive scores to the estimated log cost. OLS coefficients provide such marginal estimates, but they enforce linear effects. GAM provides marginal estimates and allows arbitrary curvature. We attempted to understand the patterns of fit by graphing our GAM-M12C5 fits versus the motor and cognitive scales. Because the GAM fits were almost as good as MART's, we thought this would give an accurate portrayal of the cost versus scale relationships. Those graphs are shown in Figures 3.11 through 3.17 for a representative selection of RICs. We computed and examined these graphs for fitting year 1999 and for pooled 1998 and 1999 data.

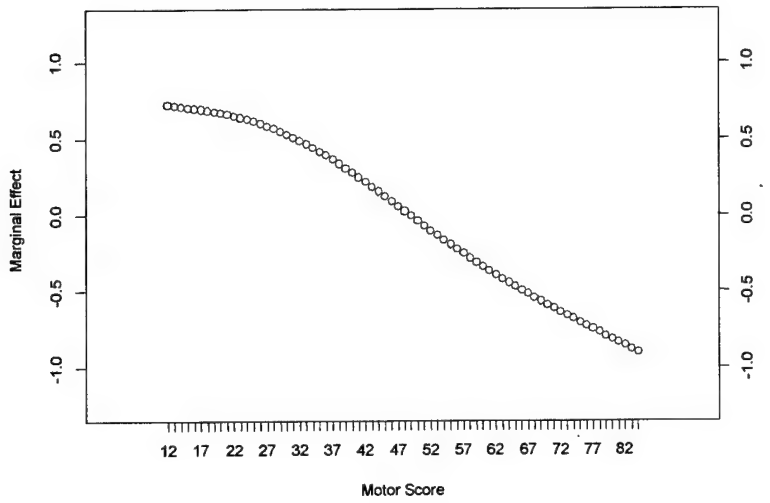

Figure 3.11—GAM Motor Scale Fits: RIC=01, Fityear=99

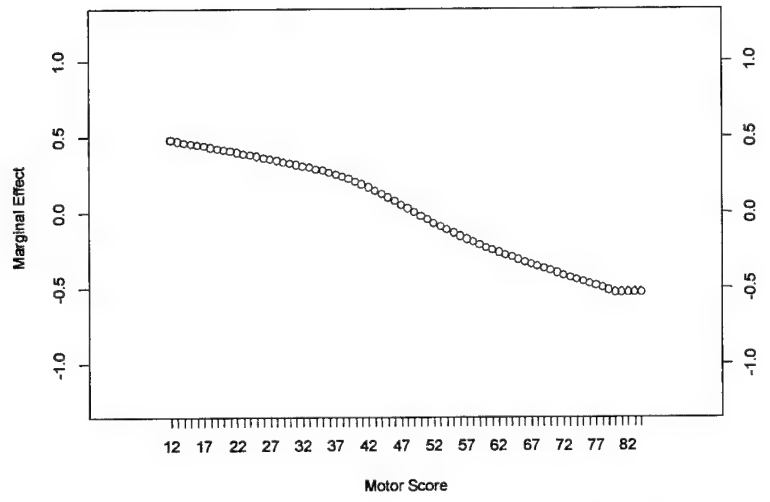

Figure 3.12—GAM Motor Scale Fits: RIC=08, Fityear=99

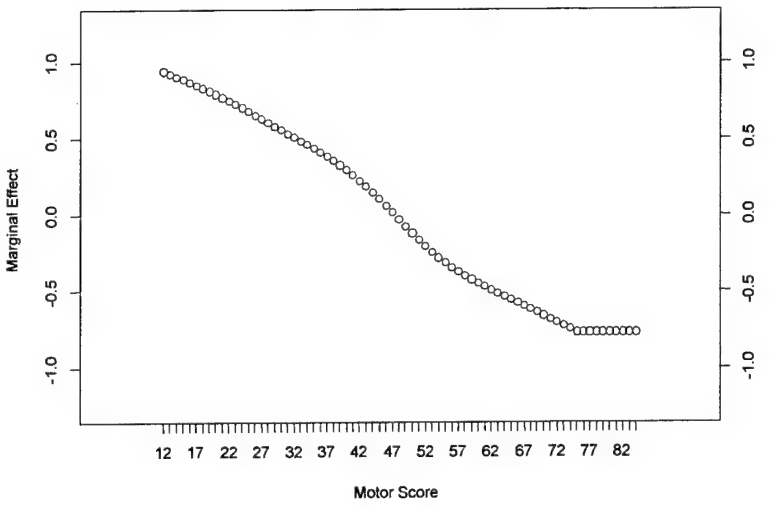

Figure 3.13—GAM Motor Scale Fits: RIC=19, Fityear=98,99

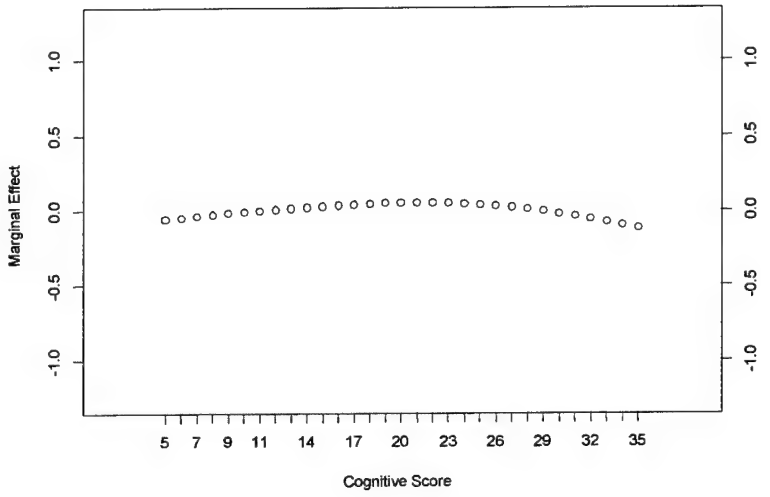

Figure 3.14—GAM Cognitive Scale Fits: RIC=01, Fityear=99

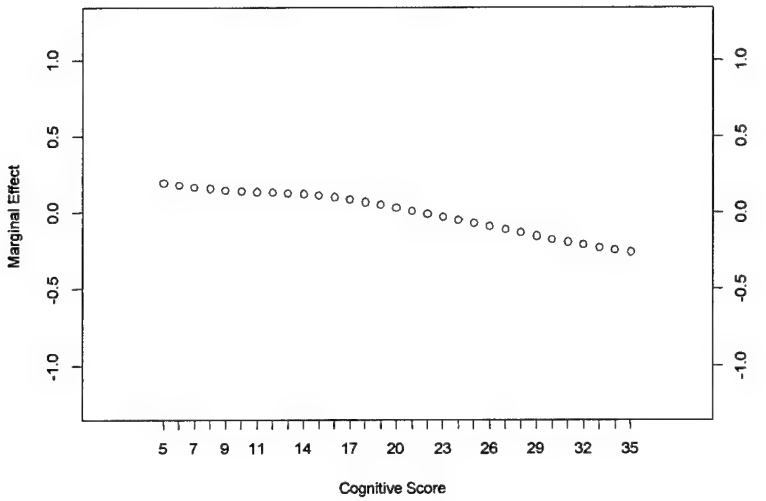

Figure 3.15—GAM Cognitive Scale Fits: RIC=02, Fityear=99

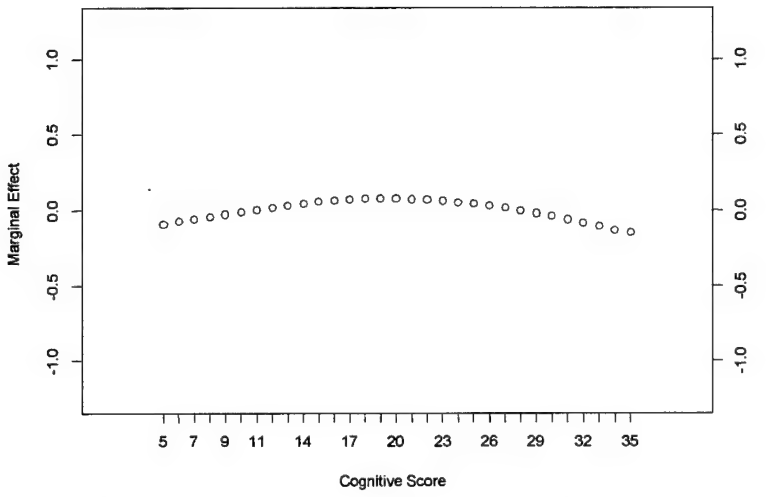

Figure 3.16—GAM Cognitive Scale Fits: RIC=08, Fityear=99

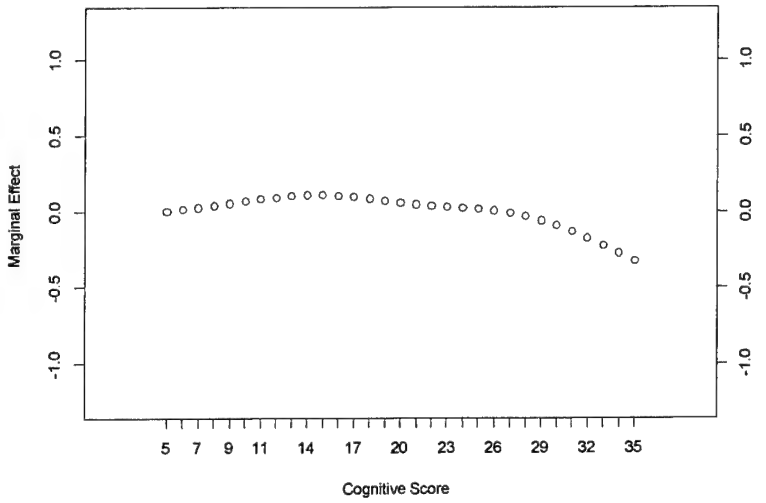

Figure 3.17—GAM Cognitive Scale Fits: RIC=18, Fityear=98,99

The plots are centered at zero, and are uniformly scaled to span the range of effects for all RICs. You can see that the motor effects are always strong and sloping in the expected direction (larger scores yield lower costs). The cognitive effects tend to be much smaller--very close to zero; also, higher scores are often associated with higher costs. Since CART attempts to replicate these patterns, it will largely split on motor scores, and hence the FRGs will simply reflect the motor score effect.

The GAM curves show cost as function of both the motor and cognitive scales. If there were discontinuities in these curves, you would expect CART to discover them and to explain a lot of the variation. But the cost curves are continuous. At best, you are asking CART to approximate a smooth curve by a (small) series of discrete jumps. This might lead you to expect a certain instability in CART's choice of cut-points, and that different data sets will indeed produce trees where cut-points differ. On the other hand, CART will find the steepness with respect to motor scores and should be expected to produce a lot of motor score splits.

Cost is strongly influenced by motor scores and in the expected direction. Except at certain motor score extremes, where there isn't much data, the higher the motor score, the lower the cost. On the other hand, the cognitive effects (Figures 3.14-3.17) are relatively flat and

frequently not monotone. This is true for indices M12C4 as well as M12C5. If we were to ask CART to discover the cost pattern, it might produce FRGs that are not monotone, which we think could pose problems for a payment system. The appropriate public policy decision might be to never lower payments for patients admitted with lower functionality--i.e., develop monotone cost curve estimates.

To see how far some of the GAM models were from monotone fits, we tried fitting the closest monotone function to each of the GAM patterns in a least-squares sense, and seeing how much of a difference this made in the percent of standard deviation explained. The results are shown in Table 3.7. The RMSEs increase by .001, a very small change. We take this as evidence that monotone cost curves fit the data almost as well as unconstrained cost curves.

Table 3.7

Change in Root Mean Squared Errors Induced by Forcing Monotone Fits

Fit Yr	Eval Yr	Regular M12C5-GAM	Monotone M12C5-GAM	Increase
99	96	469	469	.000
	97	474	475	.001
	98	479	479	.000
98-99	96	468	469	.001
	97	474	474	.000

3.3.5 Summary

Among the indices we examined, the standard cognitive index and a motor index that excludes transfer to tub do as well or better than the alternatives we examined, although a cognitive index which excludes the comprehension score also performs well and deserves further consideration.

Our current recommendation is to use M12C5 with 1SD-CART, which achieves more than 80 percent of the maximum possible reduction in standard deviation. We evaluated RMSEs and found that M12C5 outperformed the standard motor and cognitive scales. It outperformed the expanded index set lists (StJe3, StJe5) in CART models where we constrained the number of nodes, but it was about equal and slightly worse than M12C4 in its overall performance.

We find that the patterns of variation are described by a strong relationship between motor and cost--higher motor scores lower cost, and a weak relationship between cognitive and cost. The fitted curves do not appear to be far from monotone approximations that enforce an inverse relationship between cost and FIM scores. This implies that that data will support a "monotone" payment scheme where higher FIM scores never lead to higher payments, perhaps a politically desirable situation. We hope to get opinions from the TEP on this matter.

4. OBTAINING FRGS

The preceding section supports our intention to develop FRGs through 1SD-CART models using the index M12C5. In obtaining FRGs, we wanted to accommodate the following considerations.

- If it looked like we could improve the fits by the addition of a node or two, do so.
- Fits should be monotone decreasing in both the cognitive and motor indices. That is, adding a point to the FIM score should not result in a prediction that you would cost more.
- The number of nodes should be manageable--say, roughly 100. For administrative simplicity, we did not wish to create a large number of groups. In addition, we did not want to create groups characterized by very small intervals of motor or cognitive scales for fear it would encourage upcoding.
- Groups which differ on a single factor (i.e., adjacent nodes of a tree) should differ "significantly" in payment amount.

In addition, we wanted to implement these in a formal algorithm that could be judged on its own merits devoid of subjective considerations. After considerable experimentation, we arrived at the following sets of rules.

- Pool the data for 1998-99 where RIC sample sizes are less than 1,000: this results in pooling for RICs 04, 11, 18, 19, and 21. Use 1999 data only for the other RICs.
- Fit the 1SD-CART tree within each RIC.
- Consider adding nodes to the 1SD trees where R2 (equivalently, RMSE) will improve significantly. Look at traces of R2 versus number of nodes, and look to see where the addition of a node would improve R2 by 4 percent or more; add that node, and repeat this step until additional node contributions are found to be less than 4 percent.
- Produce tables describing trees and attributes of nodes (FRG numbers, N, fitted values (in dollars); flag cases where

increasing FIM scores result in higher predicted payment. These may lead to politically unacceptable payment formulas.

- Because there tends to be a monotone decreasing relationship between cost and FIM scores, at least where effects are strong, non-monotonicities tend to occur at the bottom of trees. Joining adjacent bottom nodes of a tree can eliminate these. These occur most frequently when CART is inconsistent in its splits. Consider the case where we have a series of splits on small motor score intervals and then only one of the intervals is split on age or cognitive function. This last split may introduce a discontinuity even when the underlying function is monotone. The higher cognitive function group might cost less than the total group with a lower motor score or the lower cognitive function group might cost more than the total group with a lower motor score. This kind of non-monotonicity is an artifact of CART rather than the result of actual cost patterns. We prune trees to eliminate all non-monotonicities.
- Perform additional pruning on adjacent nodes where fitted values are close (i.e., $1500). Repeat this step so long as adjacent nodes are within $1500, but do not join any nodes that would result in a predicted value that differs from the original by more than $1,000.

Table 4.1 shows how the number of nodes varied at each stage of this process. 1SD-CART started with 126 total nodes and expanded to 136 with the addition of nodes that boosted R-squared. The monotonicity requirement pared that down further the 118, and the pruning for close cost outcomes reduced that to 95. Table 4.2 shows the aggregate standard deviations for each step of this process: the monotonicity requirement and subsequent pruning affected RMSEs minimally--less than .002. Table 4.3 displays the 136-node model prior to the pruning, and shows what was grouped to accommodate both monotonicity and proximity of adjacent nodes. For example, grouping (d) was made for monotonicity purposes. Grouping (f) joined the first four lines to remove monotonicity violations, and then later added a fifth line because of node value proximity. Table 4.4 displays the final 95-node model, shown

in a format similar to Table 3.12 of the May 2000 interim report (DRU-2309-HCFA).

Table 4.1

Number of Nodes at Various Stages of Pruning

RIC	1SD	ADD	MON	FNL	Description
01	18	18	18	14	Stroke
02	5	5	5	5	Brain dysfunction, traumatic
03	4	4	4	4	Brain dysfunction, nontraumatic
04	5	5	4	4	Spinal cord dysfunction, traumatic
05	6	6	6	5	Spinal cord dysfunction, nontraumatic
06	4	4	4	4	Neurological conditions
07	16	16	10	5	Orthopedic--lower extremity fracture
08	22	22	12	6	Orthopedic--lower extremity joint repl
09	6	6	6	4	Orthopedic other
10	3	5	5	5	Amputation, lower extremity
11	2	3	3	3	Amputation, other
12	4	6	6	5	Osteoarthritis
13	3	4	4	4	Rheumatoid and other arthritis
14	4	4	4	4	Cardiac
15	4	4	4	4	Pulmonary
16	4	4	3	2	Pain syndrome
17	3	3	3	3	Major mult trauma, wo/inj to brain or spinal cord
18	2	4	4	4	Major mult trauma, w/inj to brain or spinal cord
19	2	3	3	3	Guillain-Barre
20	8	8	8	5	Other disabling impairments
21	1	2	2	2	Burns
Total	126	136	118	95	

Notes: 1SD = 1SD-CART; ADD = 1SD, plus nodes that increase R-squared; MON = ADD, after pruning for non-monotonicities; FNL = MON, after pruning where exponentiated averages are "close"

Table 4.2

RMSEs at Various Stages of Pruning

Fit Yr	Eval Yr	1SD	ADD	MON	FNL
99	96	0.474	0.474	0.474	0.475
	97	0.479	0.479	0.479	0.480
	98	0.483	0.483	0.483	0.485
	99	0.484	0.483	0.484	0.486

Notes: 1SD = 1SD-CART; ADD = 1SD, plus nodes that increase R-squared; MON = ADD, after pruning for non-monotonicities; FNL = MON, after pruning where exponentiated averages are "close"

Table 4.3

136-Node FRG Models, Before Correcting for Non-monotonicities and Proximity

RIC	FRG	N	Cost	Grouping	Condition
01	18	4215	20869		M<41.5 & M<33.5 & A<81.5 & M<26.5
	17	3763	18233		M<41.5 & M<33.5 & A<81.5 & M>26.5
	16	1065	18546		M<41.5 & M<33.5 & A>81.5 & A<88.5 & M<26.5
	15	1003	16252		M<41.5 & M<33.5 & A>81.5 & A<88.5 & M>26.5
	14	584	14750		M<41.5 & M<33.5 & A>81.5 & A>88.5
	13	3620	15756		M<41.5 & M>33.5 & M<38.5 & A<82.5
	12	987	13739		M<41.5 & M>33.5 & M<38.5 & A>82.5
	11	3102	13616		M<41.5 & M>33.5 & M>38.5
	10	4663	12149	(a)	M>41.5 & M<52.5 & M<46.5 & C<31.5
	09	1090	11037	(a)	M>41.5 & M<52.5 & M<46.5 & C>31.5
	08	2838	10754	(b)	M>41.5 & M<52.5 & M>46.5 & C<28.5 & M<50.5
	07	1254	9779	(b)	M>41.5 & M<52.5 & M>46.5 & C<28.5 & M>50.5
	06	2628	9265	(b)	M>41.5 & M<52.5 & M>46.5 & C>28.5
	05	2413	8822	(c)	M>41.5 & M>52.5 & M<58.5 & C<29.5
	04	1718	7325	(c)	M>41.5 & M>52.5 & M<58.5 & C>29.5
	03	551	7927		M>41.5 & M>52.5 & M>58.5 & C<22.5
	02	1596	6400		M>41.5 & M>52.5 & M>58.5 & C>22.5 & M<68.5
	01	250	5064		M>41.5 & M>52.5 & M>58.5 & C>22.5 & M>68.5
02	05	428	19149		M<39.5 & M<29.5
	04	400	14101		M<39.5 & M>29.5
	03	602	11522		M>39.5 & C<23.5
	02	303	9858		M>39.5 & C>23.5 & M<51.5
	01	320	7137		M>39.5 & C>23.5 & M>51.5
03	04	442	20333		M<40.5 & M<24.5
	03	1099	14429		M<40.5 & M>24.5
	02	1164	10754		M>40.5 & M<50.5
	01	1053	8168		M>40.5 & M>50.5
04	05	111	14913	(d)	M<35.5 & M<18.5 & A<55.5
	04	171	26635	(d)	M<35.5 & M<18.5 & A>55.5
	03	604	16236		M<35.5 & M>18.5
	02	599	11282		M>35.5 & M<49.5
	01	398	7785		M>35.5 & M>49.5
05	06	1139	16882		M<40.5 & M<33.5
	05	923	11837		M<40.5 & M>33.5
	04	955	9321	(e)	M>40.5 & M<50.5 & M<45.5
	03	1079	8063	(e)	M>40.5 & M<50.5 & M>45.5
	02	243	7951		M>40.5 & M>50.5 & C<29.5
	01	1498	6317		M>40.5 & M>50.5 & C>29.5
06	04	2221	13373		M<46.5 & M<35.5
	03	2924	10982		M<46.5 & M>35.5
	02	2486	8911		M>46.5 & M<55.5
	01	1244	6988		M>46.5 & M>55.5

Table 4.3 (cont.)

07	16	374	11861	(f)	M<45.5 & M<37.5 & M<33.5 & C<13.5
	15	2330	13602	(f)	M<45.5 & M<37.5 & M<33.5 & C>13.5 & C<33.5
	14	149	10007	(f)	M<45.5 & M<37.5 & M<33.5 & C>13.5 & C>33.5 & C<34.5
	13	355	13135	(f)	M<45.5 & M<37.5 & M<33.5 & C>13.5 & C>33.5 & C>34.5
	12	2189	11885	(f)	M<45.5 & M<37.5 & M>33.5
	11	1471	11603	(g)	M<45.5 & M>37.5 & M<41.5 & C<30.5
	10	1562	10583	(g)	M<45.5 & M>37.5 & M<41.5 & C>30.5
	09	706	10394	(h)	M<45.5 & M>37.5 & M>41.5 & A<81.5 & C<30.5
	08	1264	9284	(h)	M<45.5 & M>37.5 & M>41.5 & A<81.5 & C>30.5
	07	1597	10530	(h)	M<45.5 & M>37.5 & M>41.5 & A>81.5
	06	1995	9274	(i)	M>45.5 & M<51.5 & C<31.5
	05	1551	8665	(i)	M>45.5 & M<51.5 & C>31.5 & M<48.5
	04	1593	7919	(i)	M>45.5 & M<51.5 & C>31.5 & M>48.5
	03	349	8787	(j)	M>45.5 & M>51.5 & M<55.5 & C<29.5
	02	1792	7245	(j)	M>45.5 & M>51.5 & M<55.5 & C>29.5
	01	1350	6380	(j)	M>45.5 & M>51.5 & M>55.5
08	22	1411	11237	(k)	M<46.5 & C<31.5 & M<39.5 & M<34.5
	21	1323	10007	(k)	M<46.5 & C<31.5 & M<39.5 & M>34.5
	20	3083	8544		M<46.5 & C<31.5 & M>39.5
	19	749	8647	(l)	M<46.5 & C>31.5 & M<42.5 & A<80.5 & C<34.5 & C<33.5
	18	408	5603	(l)	M<46.5 & C>31.5 & M<42.5 & A<80.5 & C<34.5 & C>33.5 & M<36.5
	17	404	7879	(l)	M<46.5 & C>31.5 & M<42.5 & A<80.5 & C<34.5 & C>33.5 & M>36.5
	16	499	9576	(l)	M<46.5 & C>31.5 & M<42.5 & A<80.5 & C>34.5 & M<35.5
	15	1607	7927	(l)	M<46.5 & C>31.5 & M<42.5 & A<80.5 & C>34.5 & M>35.5
	14	1095	9100	(l)	M<46.5 & C>31.5 & M<42.5 & A>80.5
	13	2153	6898	(l)	M<46.5 & C>31.5 & M>42.5 & A<74.5
	12	2408	7518	(l)	M<46.5 & C>31.5 & M>42.5 & A>74.5
	11	2674	7023	(m)	M>46.5 & M<54.5 & C<33.5 & A<81.5 & M<50.5
	10	2259	6400	(m)	M>46.5 & M<54.5 & C<33.5 & A<81.5 & M>50.5
	09	1378	7548	(m)	M>46.5 & M<54.5 & C<33.5 & A>81.5
	08	3594	6438	(m)	M>46.5 & M<54.5 & C>33.5 & M<49.5 & A<81.5
	07	623	7274	(m)	M>46.5 & M<54.5 & C>33.5 & M<49.5 & A>81.5
	06	5406	5779	(m)	M>46.5 & M<54.5 & C>33.5 & M>49.5 & A<77.5
	05	2353	6260	(m)	M>46.5 & M<54.5 & C>33.5 & M>49.5 & A>77.5
	04	653	6342	(n)	M>46.5 & M>54.5 & M<57.5 & C<31.5
	03	4186	5443	(n)	M>46.5 & M>54.5 & M<57.5 & C>31.5
	02	3839	5162	(o)	M>46.5 & M>54.5 & M>57.5 & M<62.5
	01	1322	4666	(o)	M>46.5 & M>54.5 & M>57.5 & M>62.5
09	06	1777	11920		M<46.5 & M<37.5

Table 4.3 (cont.)

	05	1202	10148	(p)	M<46.5 & M>37.5 & C<30.5
	04	1713	8822	(p)	M<46.5 & M>37.5 & C>30.5
	03	2905	7692		M>46.5 & M<53.5
	02	432	6857	(q)	M>46.5 & M>53.5 & C<31.5
	01	1281	5722	(q)	M>46.5 & M>53.5 & C>31.5
10	05	1495	14794		M<45.5 & M<38.5
	04	1319	12531		M<45.5 & M>38.5
	03	1465	10938		M>45.5 & M<51.5
	02	1412	9377		M>45.5 & M>51.5 & M<60.5
	01	465	7809		M>45.5 & M>51.5 & M>60.5
11	03	217	14300		M<51.5 & M<37.5
	02	580	10711		M<51.5 & M>37.5
	01	407	7793		M>51.5
12	06	861	12772		M<47.5 & M<38.5
	05	1403	10342		M<47.5 & M>38.5
	04	862	8920	(r)	M>47.5 & M<54.5 & C<33.5
	03	781	7785	(r)	M>47.5 & M<54.5 & C>33.5
	02	489	7578		M>47.5 & M>54.5 & C<33.5
	01	640	6027		M>47.5 & M>54.5 & C>33.5
13	04	397	13427		M<46.5 & M<35.5
	03	710	10097		M<46.5 & M>35.5
	02	660	8358		M>46.5 & M<53.5
	01	583	6667		M>46.5 & M>53.5
14	04	1018	12657		M<47.5 & M<37.5
	03	2200	9887		M<47.5 & M>37.5
	02	2747	7871		M>47.5 & M<55.5
	01	2139	6298		M>47.5 & M>55.5
15	04	629	15460		M<47.5 & M<35.5
	03	1367	11328		M<47.5 & M>35.5
	02	2461	9072		M>47.5 & M<60.5
	01	925	7662		M>47.5 & M>60.5
16	04	958	9789		M<44.5
	03	981	7856	(s)	M>44.5 & M<52.5
	02	166	8656	(s)	M>44.5 & M>52.5 & A<63.5
	01	888	6323	(s)	M>44.5 & M>52.5 & A>63.5
17	03	333	15139		M<45.5 & M<32.5
	02	750	11339		M<45.5 & M>32.5
	01	596	8136		M>45.5
18	04	87	25336		M<44.5 & M<25.5
	03	229	14516		M<44.5 & M>25.5
	02	103	9927		M>44.5 & C<32.5
	01	58	6470		M>44.5 & C>32.5
19	03	143	25591		M<46.5 & M<30.5
	02	245	16916		M<46.5 & M>30.5
	01	224	9274		M>46.5
20	08	1812	14415		M<44.5 & M<32.5 & A<81.5

Table 4.3 (cont.)

	07	854	12173		M<44.5 & M<32.5 & A>81.5
	06	2134	11861	(t)	M<44.5 & M>32.5 & M<37.5
	05	4405	10530	(t)	M<44.5 & M>32.5 & M>37.5
	04	4305	9145	(u)	M>44.5 & M<53.5 & M<49.5
	03	3322	8259	(u)	M>44.5 & M<53.5 & M>49.5
	02	3132	7347	(v)	M>44.5 & M>53.5 & M<59.5
	01	1589	6400	(v)	M>44.5 & M>53.5 & M>59.5
21	02	119	17677		M<45.5
	01	94	9284		M>45.5

Notes: M stands for the 12-component FIM motor score, C for the standard FIM cognitive score, and A for age. The FRG numbers were assigned by CART mostly in increasing order of average cost, although exceptions were made to keep adjacent nodes together.

Table 4.4

Recommended 95-Node FRG Models

RIC	FRG	N	Cost	Condition
01	14	4215	20869	M<41.5 & M<33.5 & A<81.5 & M<26.5
	13	3763	18233	M<41.5 & M<33.5 & A<81.5 & M>26.5
	12	1065	18546	M<41.5 & M<33.5 & A>81.5 & A<88.5 & M<26.5
	11	1003	16252	M<41.5 & M<33.5 & A>81.5 & A<88.5 & M>26.5
	10	584	14750	M<41.5 & M<33.5 & A>81.5 & A>88.5
	09	3620	15756	M<41.5 & M>33.5 & M<38.5 & A<82.5
	08	987	13739	M<41.5 & M>33.5 & M<38.5 & A>82.5
	07	3102	13616	M<41.5 & M>33.5 & M>38.5
	06	5753	11932	M>41.5 & M<52.5 & M<46.5
	05	6720	9967	M>41.5 & M<52.5 & M>46.5
	04	4131	8168	M>41.5 & M>52.5 & M<58.5
	03	551	7927	M>41.5 & M>52.5 & M>58.5 & C<22.5
	02	1596	6400	M>41.5 & M>52.5 & M>58.5 & C>22.5 & M<68.5
	01	250	5064	M>41.5 & M>52.5 & M>58.5 & C>22.5 & M>68.5
02	05	428	19149	M<39.5 & M<29.5
	04	400	14101	M<39.5 & M>29.5
	03	602	11522	M>39.5 & C<23.5
	02	303	9858	M>39.5 & C>23.5 & M<51.5
	01	320	7137	M>39.5 & C>23.5 & M>51.5
03	04	442	20333	M<40.5 & M<24.5
	03	1099	14429	M<40.5 & M>24.5
	02	1164	10754	M>40.5 & M<50.5
	01	1053	8168	M>40.5 & M>50.5
04	04	282	21248	M<35.5 & M<18.5
	03	604	16236	M<35.5 & M>18.5
	02	599	11282	M>35.5 & M<49.5
	01	398	7785	M>35.5 & M>49.5
05	05	1139	16882	M<40.5 & M<33.5
	04	923	11837	M<40.5 & M>33.5
	03	2034	8630	M>40.5 & M<50.5
	02	243	7951	M>40.5 & M>50.5 & C<29.5
	01	1498	6317	M>40.5 & M>50.5 & C>29.5
06	04	2221	13373	M<46.5 & M<35.5
	03	2924	10982	M<46.5 & M>35.5
	02	2486	8911	M>46.5 & M<55.5
	01	1244	6988	M>46.5 & M>55.5
07	05	5397	12620	M<45.5 & M<37.5
	04	3033	11059	M<45.5 & M>37.5 & M<41.5
	03	3567	10047	M<45.5 & M>37.5 & M>41.5
	02	5139	8656	M>45.5 & M<51.5
	01	3491	7030	M>45.5 & M>51.5
08	06	2734	10625	M<46.5 & C<31.5 & M<39.5
	05	3083	8544	M<46.5 & C<31.5 & M>39.5
	04	9323	7708	M<46.5 & C>31.5

Table 4.4 (cont.)

	03	18287	6393	M>46.5 & M<54.5
	02	4839	5552	M>46.5 & M>54.5 & M<57.5
	01	5161	5029	M>46.5 & M>54.5 & M>57.5
09	04	1777	11920	M<46.5 & M<37.5
	03	2915	9339	M<46.5 & M>37.5
	02	2905	7692	M>46.5 & M<53.5
	01	1713	5991	M>46.5 & M>53.5
10	05	1495	14794	M<45.5 & M<38.5
	04	1319	12531	M<45.5 & M>38.5
	03	1465	10938	M>45.5 & M<51.5
	02	1412	9377	M>45.5 & M>51.5 & M<60.5
	01	465	7809	M>45.5 & M>51.5 & M>60.5
11	03	217	14300	M<51.5 & M<37.5
	02	580	10711	M<51.5 & M>37.5
	01	407	7793	M>51.5
12	05	861	12772	M<47.5 & M<38.5
	04	1403	10342	M<47.5 & M>38.5
	03	1643	8358	M>47.5 & M<54.5
	02	489	7578	M>47.5 & M>54.5 & C<33.5
	01	640	6027	M>47.5 & M>54.5 & C>33.5
13	04	397	13427	M<46.5 & M<35.5
	03	710	10097	M<46.5 & M>35.5
	02	660	8358	M>46.5 & M<53.5
	01	583	6667	M>46.5 & M>53.5
14	04	1018	12657	M<47.5 & M<37.5
	03	2200	9887	M<47.5 & M>37.5
	02	2747	7871	M>47.5 & M<55.5
	01	2139	6298	M>47.5 & M>55.5
15	04	629	15460	M<47.5 & M<35.5
	03	1367	11328	M<47.5 & M>35.5
	02	2461	9072	M>47.5 & M<60.5
	01	925	7662	M>47.5 & M>60.5
16	02	958	9789	M<44.5
	01	2035	7201	M>44.5
17	03	333	15139	M<45.5 & M<32.5
	02	750	11339	M<45.5 & M>32.5
	01	596	8136	M>45.5
18	04	87	25336	M<44.5 & M<25.5
	03	229	14516	M<44.5 & M>25.5
	02	103	9927	M>44.5 & C<32.5
	01	58	6470	M>44.5 & C>32.5
19	03	143	25591	M<46.5 & M<30.5
	02	245	16916	M<46.5 & M>30.5
	01	224	9274	M>46.5
20	05	1812	14415	M<44.5 & M<32.5 & A<81.5
	04	854	12173	M<44.5 & M<32.5 & A>81.5
	03	6539	10949	M<44.5 & M>32.5
	02	7627	8752	M>44.5 & M<53.5

	01	4721	7016	M>44.5 & M>53.5
21	02	119	17677	M<45.5
	01	94	9284	M>45.5

Notes: M stands for the 12-component FIM motor score, C for the standard FIM cognitive score, and A for age. The FRG numbers were assigned by CART mostly in increasing order of average cost, although exceptions were made to keep adjacent nodes together.

The final trees differ in some respects from the trees produced in the interim report. This is not surprising--CART is trying to fit step functions to continuous curves, so the cut-points are imprecisely determined. We think the important question is not whether the trees are identical but instead whether the tree models produce a consistent and accurate set of predictions. For now, we simply ask if one used the current FRGs and associated predictions, how different are these predictions over the different years?

The attached plots demonstrate that this set of FRG models fits the data pretty well in all years. They show the predicted means within FRGs; predictions are normalized to have the same mean as actual log-cost across all RICs. Except for the RICs that were pooled, the 1999 predicted means fall right on top of the log cost averages (rightmost panels). As you move to the left, you see the models projected backwards in time. The predictions for 1998 look quite good. They are a little worse for 1997, and a little worse still for 1996.

An upcoming part of the project will consider how much of a difference we will need to conclude that the models need to be refit, and how this refit might occur. We will gain some insight into the performance of various algorithms such as adjusting cut-points, splitting FRGs, combining FRGs, or completely refitting the CART models by using these plots to identify anomalous patterns and seeing what it takes to eliminate them. We are already thinking of making some adjustments to standardize on age group splits. For example, RIC 01 uses age<81.5 in some places, A<82.5 in others: we would like the TEP's opinion on the importance of uniformity in age cut-points.

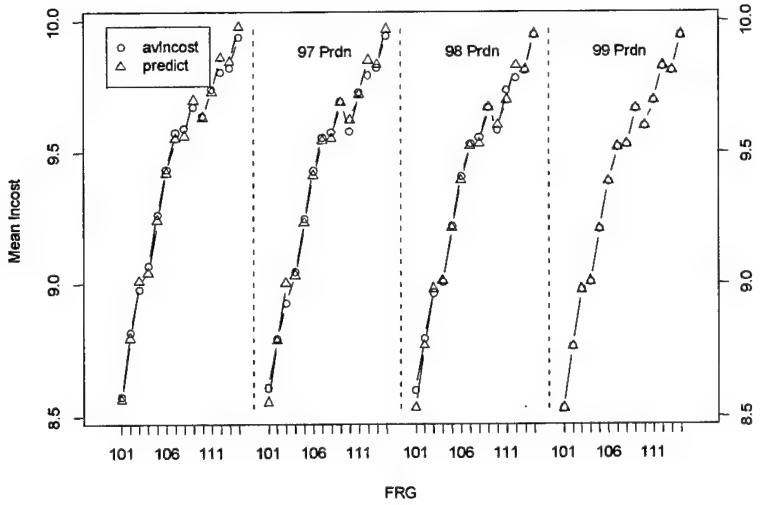

Figure 4.1—Actual and Predicted FRG Means: RIC=01, Fityear=99

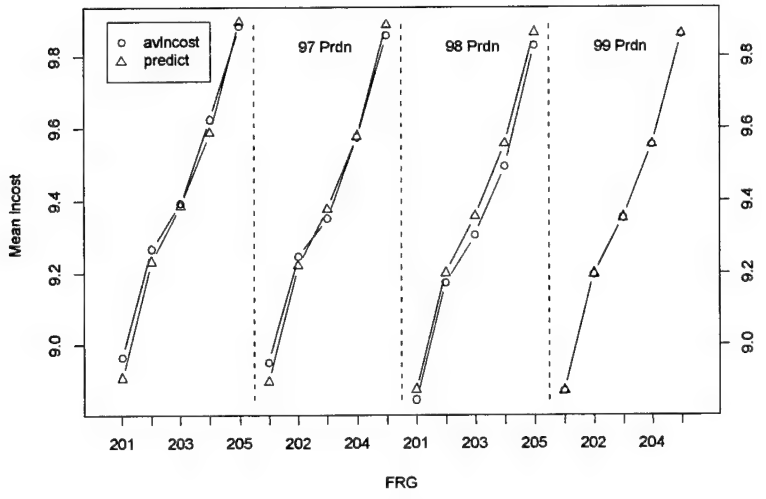

Figure 4.2—Actual and Predicted FRG Means: RIC=02, Fityear=99

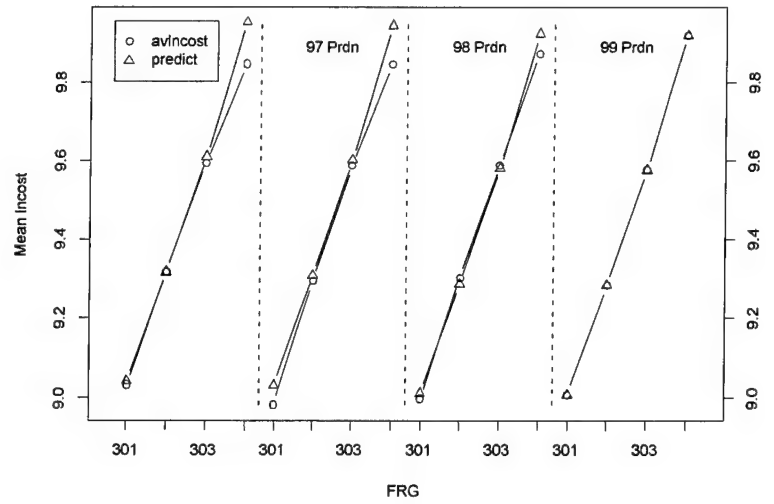

Figure 4.3—Actual and Predicted FRG Means: RIC=03, Fityear=99

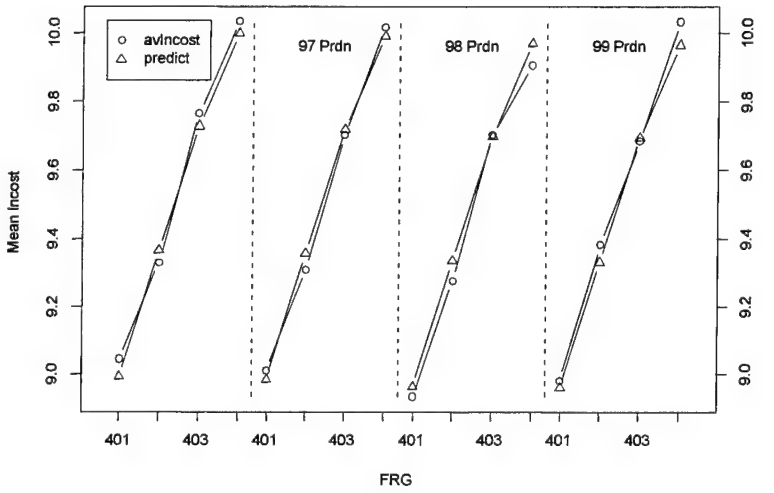

Figure 4.4—Actual and Predicted FRG Means: RIC=04, Fityear=98,99

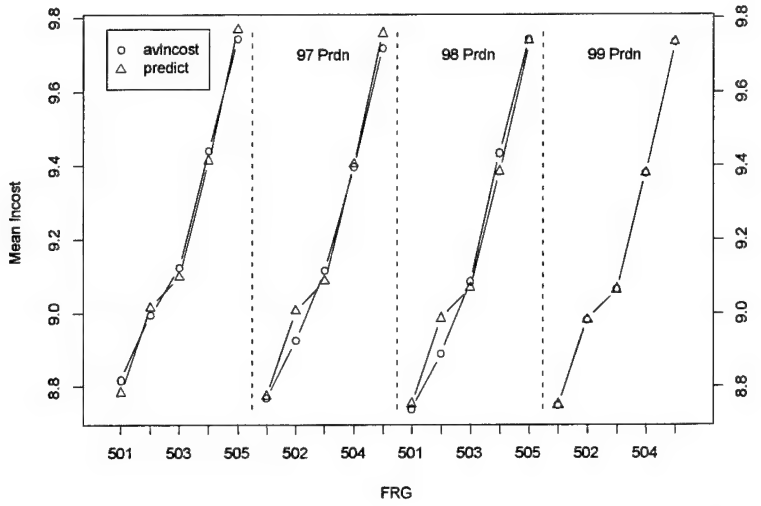

Figure 4.5—Actual and Predicted FRG Means: RIC=05, Fityear=99

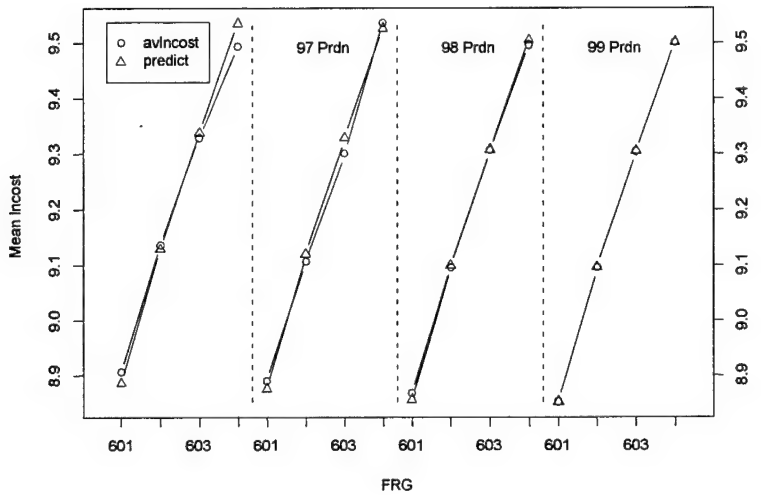

Figure 4.6—Actual and Predicted FRG Means: RIC=06, Fityear=99

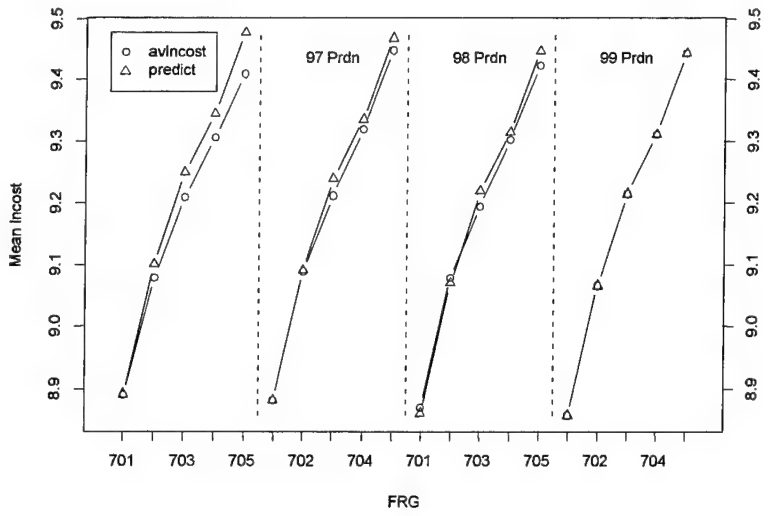

Figure 4.7—Actual and Predicted FRG Means: RIC=07, Fityear=99

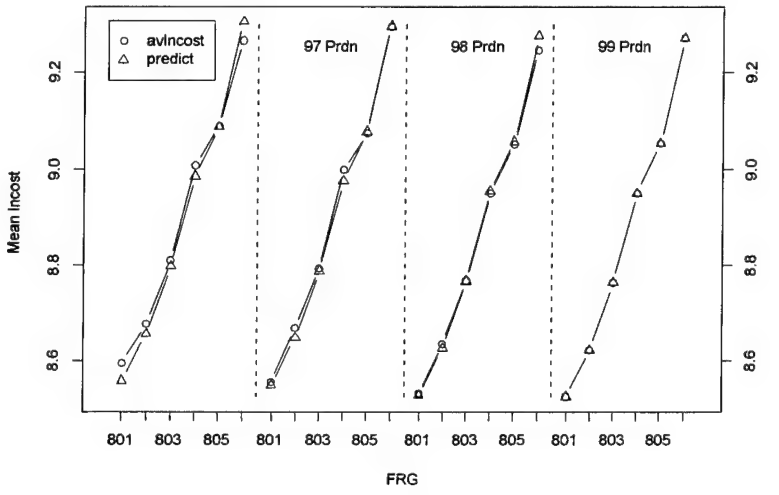

Figure 4.8—Actual and Predicted FRG Means: RIC=08, Fityear=99

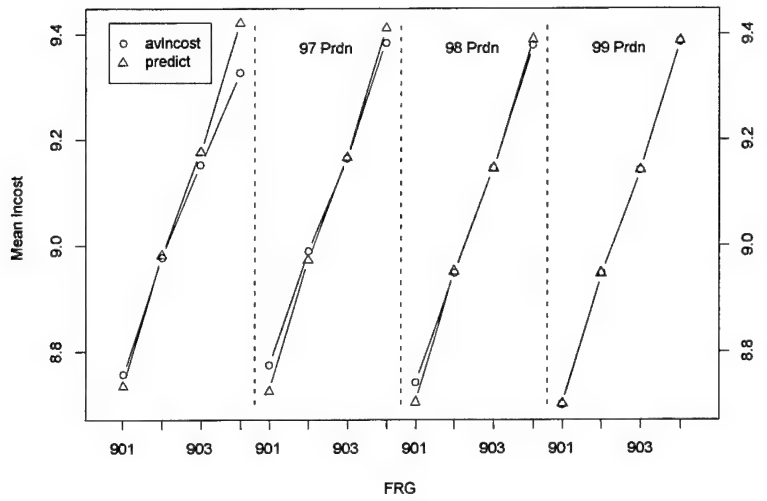

Figure 4.9—Actual and Predicted FRG Means: RIC=09, Fityear=99

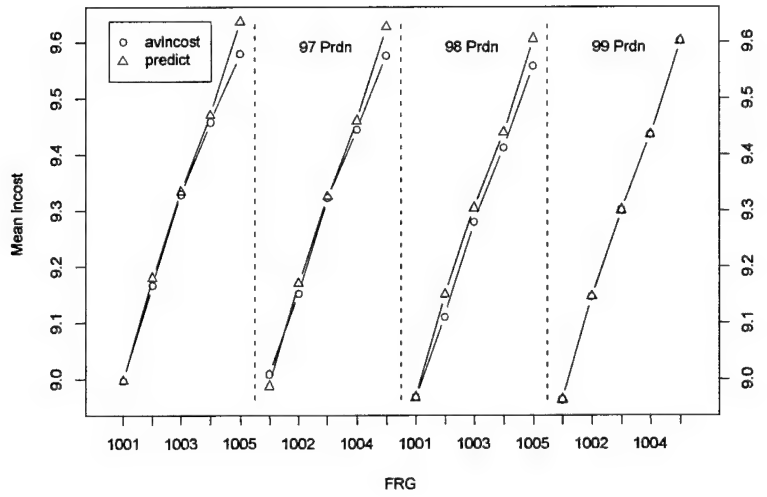

Figure 4.10—Actual and Predicted FRG Means: RIC=10, Fityear=99

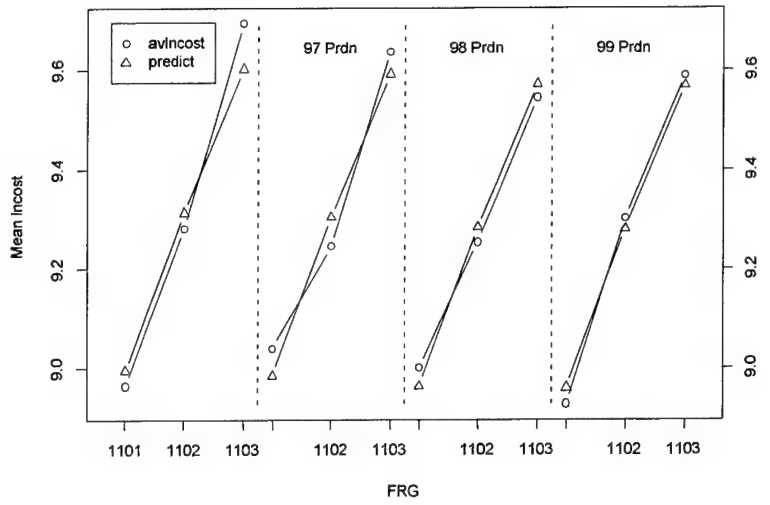

Figure 4.11—Actual and Predicted FRG Means: RIC=11, Fityear=99

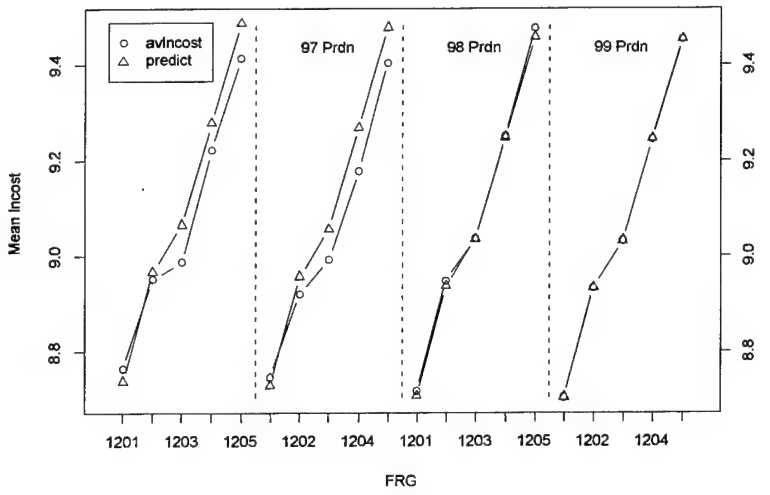

Figure 4.12—Actual and Predicted FRG Means: RIC=12, Fityear=99

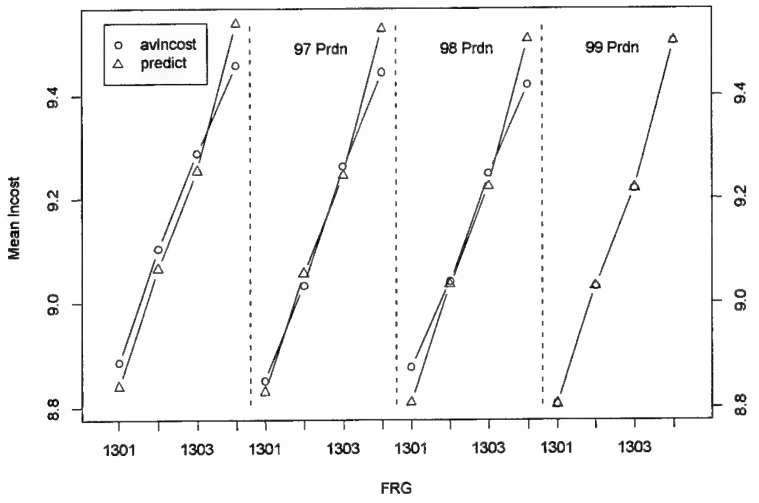

Figure 4.13—Actual and Predicted FRG Means: RIC=13, Fityear=99

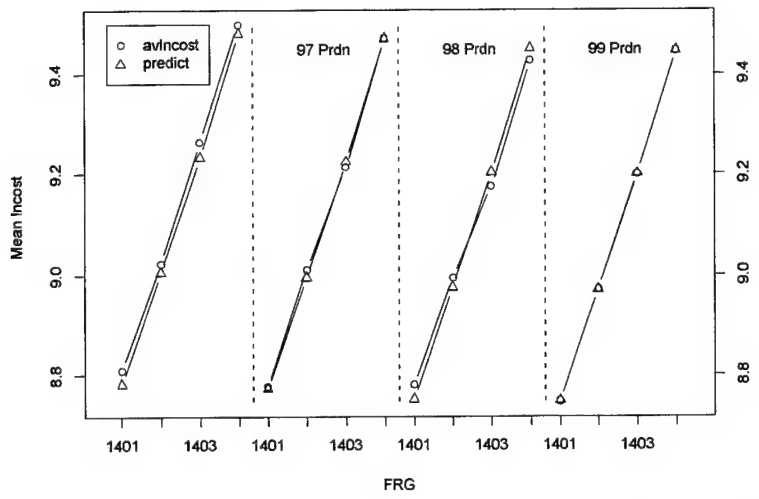

Figure 4.14—Actual and Predicted FRG Means: RIC=14, Fityear=99

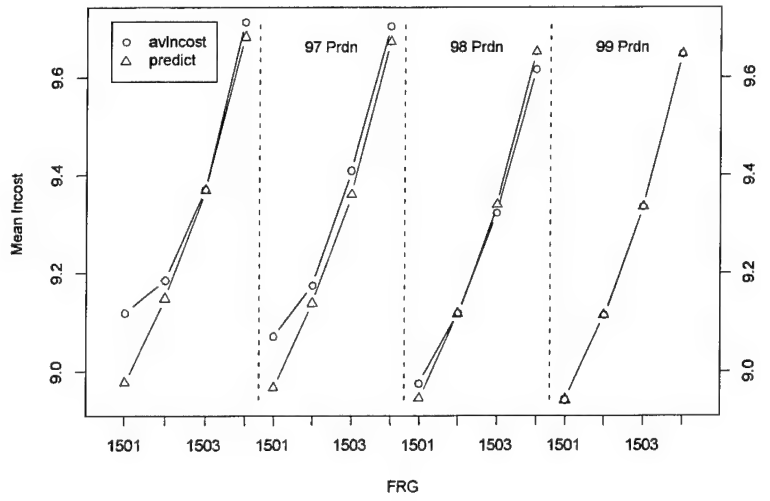

Figure 4.15—Actual and Predicted FRG Means: RIC=15, Fityear=99

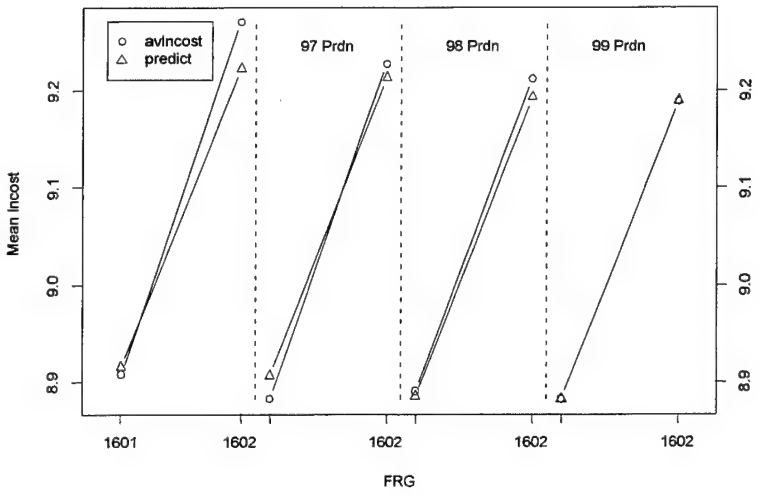

Figure 4.16—Actual and Predicted FRG Means: RIC=16, Fityear=99

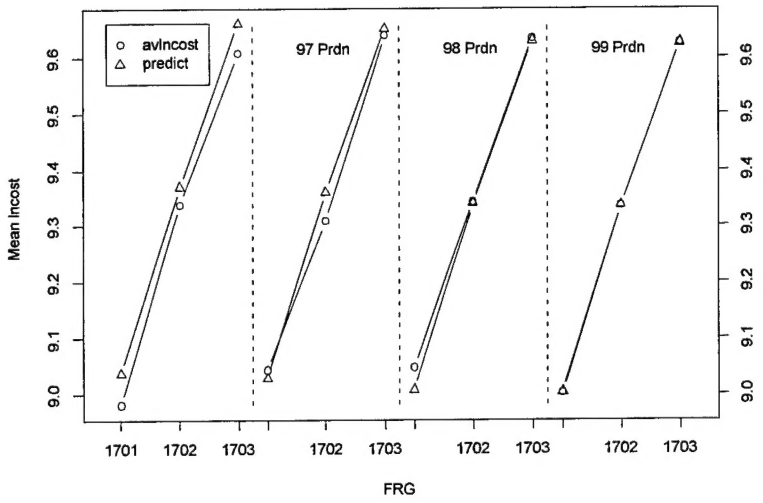

Figure 4.17—Actual and Predicted FRG Means: RIC=17, Fityear=99

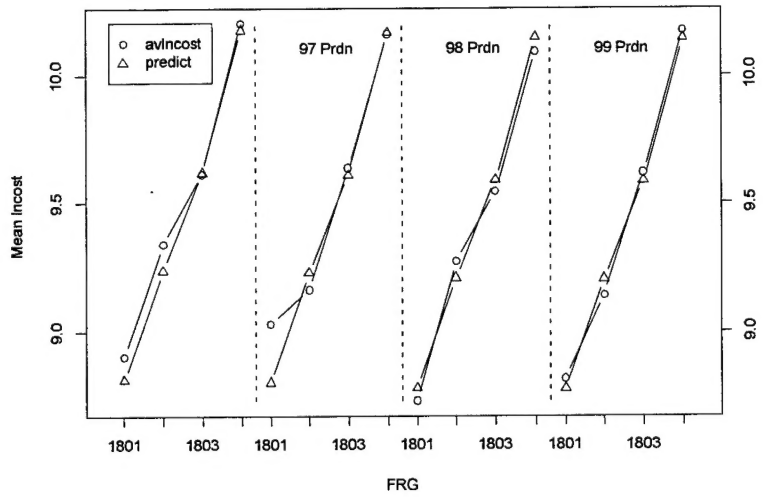

Figure 4.18—Actual and Predicted FRG Means: RIC=18, Fityear=98,99

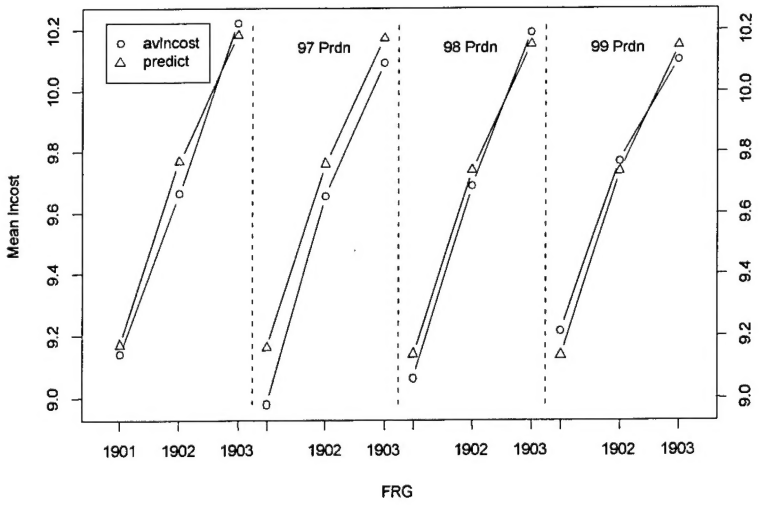

Figure 4.19—Actual and Predicted FRG Means: RIC=19, Fityear=98,99

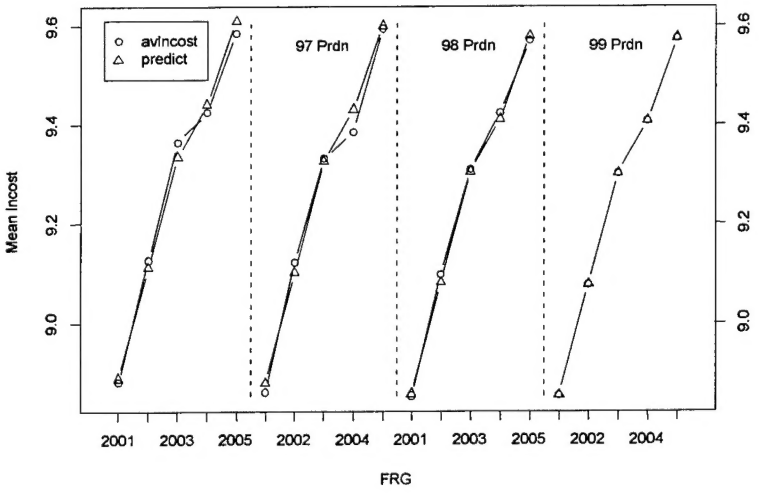

Figure 4.20—Actual and Predicted FRG Means: RIC=20, Fityear=99

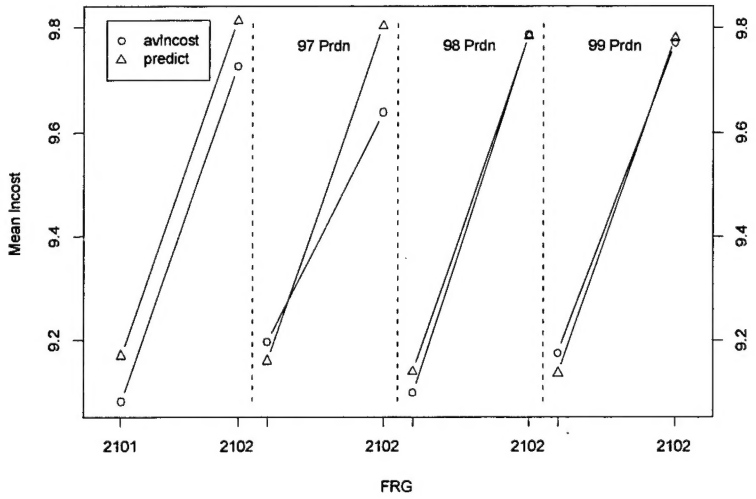

Figure 4.21—Actual and Predicted FRG Means: RIC=21, Fityear=98,99

REFERENCES

Breiman, L. J., M. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees.* Belmont, CA: Wandsworth, Inc.

Carter, Grace M., Daniel A. Relles, and Barbara O. Wynn (January 2000). *Workplan for an Inpatient Rehabilitation Prospective Payment System.* Santa Monica, CA: RAND, DRU-2161-1-HCFA.

Carter, Grace M., Daniel A. Relles, Barbara O. Wynn, Jennifer Kawata, Susan M. Paddock, Neeraj Sood, and Mark E. Totten (July 2000). *Interim Report on an Inpatient Rehabilitation Facility Prospective Payment System.* Santa Monica, CA: RAND, DRU-2309-HCFA.

Friedman, J. H., T. Hastie, and R. J. Tibshirani (April 2000). Additive Logisitc Regression: A Statistical View of Boosting. *Annals of Statistics, Vol. 28, No. 2,* pp. 337-407.

Hastie, T. and R. J. Tibshirani (1990). *Generalized Additive Models.* London: Chapman and Hall.

Stineman, M. G., A. J. Jette, R. C. Fiedler, and C. V. Granger (June 1997a). Impairment-Specific Dimensions Within the Functional Independence Measure. *Archives of Physical Medical Rehabilitation, Vol. 78,* pp. 636-643.

Stineman, M. G., C. J. Tassoni, J. J. Escarce, J. E. Goin, C. V. Granger, R. C. Fiedler, and S. V. Williams (October 1997b). Development of Function-Related Groups, Version 2.0: A Classification System for Medical Rehabilitation. *Health Services Research, Vol. 32, No. 4,* pp. 529-548.

Uniform Data System for Medical Rehabilitation (UDSmr) (1997). *Guide for the Uniform Data Set for Medical Rehabilitation, Version 5.1.* Buffalo, NY: UDSmr.